3822–3833 Nucleic Acids Research, 2012, Vol. 40, No. 9 Published online 12 January 2012
doi:10.1093/nar/gkr1295

New methods for finding common insertion sites and co-occurring common insertion sites in transposon- and virus-based genetic screens

Tracy L. Bergemann[1,*], Timothy K. Starr[2,3,*], Haoyu Yu[4], Michael Steinbach[5], Jesse Erdmann[6], Yun Chen[7], Robert T. Cormier[8], David A. Largaespada[3] and Kevin A. T. Silverstein[6]

[1]Division of Biostatistics, School of Public Health, [2]Department of Obstetrics, Gynecology & Women's Health, [3]Department of Genetics, Cell Biology and Development, Masonic Cancer Center, and Center for Genome Engineering, [4]Minnesota Supercomputing Institute, [5]Department of Computer Science and Engineering, [6]Biostatistics and Bioinformatics, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, [7]Opera Solutions, Cambridge, MA 02142 and [8]Department of Biochemistry and Molecular Biology, University of Minnesota Medical School, Duluth, MN 55812, USA

Received September 20, 2010; Revised December 11, 2011; Accepted December 15, 2011

ABSTRACT

Insertional mutagenesis screens in mice are used to identify individual genes that drive tumor formation. In these screens, candidate cancer genes are identified if their genomic location is proximal to a common insertion site (CIS) defined by high rates of transposon or retroviral insertions in a given genomic window. In this article, we describe a new method for defining CISs based on a Poisson distribution, the Poisson Regression Insertion Model, and show that this new method is an improvement over previously described methods. We also describe a modification of the method that can identify pairs and higher orders of co-occurring common insertion sites. We apply these methods to two data sets, one generated in a transposon-based screen for gastrointestinal tract cancer genes and another based on the set of retroviral insertions in the Retroviral Tagged Cancer Gene Database. We show that the new methods identify more relevant candidate genes and candidate gene pairs than found using previous methods. Identification of the biologically relevant set of mutations that occur in a single cell and cause tumor progression will aid in the rational design of single and combinatorial therapies in the upcoming age of personalized cancer therapy.

INTRODUCTION

Forty years ago, cancer researchers hypothesized, based on epidemiological analysis of cancer rates that tumors develop after the stochastic acquisition of multiple somatic mutations (1). Although there are a few cancers whose etiology can be assigned to one or two mutations, such as chronic myeloid leukemia and retinoblastoma, the majority of cancers are likely the result of many mutations (2,3). Furthermore, a series of tumorigenesis stages have been described with the suggestion that each stage follows from the acquisition of one or more new driver mutations and that disease severity may be linked with the number of driver mutations (4,5). To better understand the genetic basis of tumorigenesis in order to develop improved therapies, biological and mathematical models must be developed that can adequately describe the complex interaction of multiple genetic events that cause a cancer phenotype.

To address this need, we have developed a powerful system to model multi-hit tumorigenesis in mice using a sleeping beauty (SB) transposon-based forward genetic screen (6–10). This method is similar to retroviral mutagenesis screens that have been used to successfully identify cancer genes in hematopoietic, mammary and brain tumors (11). Both methods rely on gain- and loss-of-function mutations generated by random transposon or provirus insertions. When a single cell accumulates the correct combination of mutations, it proliferates and

*To whom correspondence should be addressed. Tel: +612 625 9142; Fax: +612 626 0660; Email: tracy.l.bergemann@medtronic.com
Correspondence may also be addressed to Timothy K. Starr. Tel: +612 625 4425; Fax: +612 626 0665; Email: star0044@umn.edu
Present address:
Tracy L. Bergemann, Cardiac Rhythm Disease Management, Medtronic, Mounds View, MN 55112, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

becomes a tumor. The hypothesis is that the transposon or provirus provides several key mutations, although not necessarily all of the mutations required for tumor initiation and progression. Finding the genomic location of the insertions in the tumor DNA is relatively easy using specialized PCR techniques combined with high-throughput sequencing and BLAST analysis. In this manner, it is possible to map several hundred insertions in a single tumor. The majority of these mapped insertions, however, does not contribute to the tumor phenotype and are referred to as passenger mutations, while only a small subset of insertions generated the driver mutations that caused the cancer. To identify these driver mutations, a statistical approach is used to find common insertion sites (CISs), defined as genomic loci experiencing insertions at a rate higher than expected by chance. The idea being that it would be a rare occurrence for multiple tumors to harbor an insertion in the same locus by chance unless that insertion contributed to tumor formation. The genes within these CISs are then considered candidate cancer genes based on the assumption that the transposon or provirus created a mutation in these genes that contributed to tumor growth.

Previous statistical approaches used to identify CISs have been based on the Poisson distribution (12), Monte Carlo simulations (9,10) and kernel convolution (13). It has become increasingly evident that modifications to these approaches are necessary as the size of the screens increase and the biology of transposon and provirus insertion site preferences are better understood. For example, Monte Carlo simulations become computationally expensive as the size of the screen increases. Thus, larger data sets will require a model-based approach that provides the same inference as Monte Carlo simulations. Additionally, none of the methods, when applied evenly across the genome, are able to account for the uneven distribution of site preferences, such as the distribution of TA dinucleotides, which are the required insertion sites for some transposons, or the distribution of transcriptional start sites (TSSs), which are the preferred sites of viral integration. Appropriate models for tumorigenesis will also require a statistical method that can identify pairs or higher orders of co-occurring mutations in a single tumor in order to find cooperating mutations that cause cancer. These higher orders of insertions in a single tumor are called common co-occurring insertions (CCIs) to distinguish them from CISs. In an attempt to solve some of these problems, de Ridder *et al.* (13,14) introduced the Gaussian kernel convolution framework as a method to identify CISs and CCIs. The framework can be scaled to adjust for genomic window size and can adjust for TSS bias. Unfortunately, similar to Monte–Carlo simulations, the CCI detection method becomes unwieldy when analyzing larger data sets because it relies on a permutation strategy for inference that currently requires hundreds to thousands of CPU hours. We also find that their CCI detection is seriously affected by the subjective choice of permutation strategy. Further, the Gaussian kernel convolution, which smoothes both the insertion counts and the distribution of TSSs, assumes a continuous process to approximate underlying count data which can lead to bias in the sensitivity and specificity of detection.

In this article, we present a new model to define CISs and CCIs, which we refer to as the PRIM method. Because the Poisson distribution describes count data, this model-based approach accurately reflects Monte–Carlo simulations of the same insertion process. We describe how to directly incorporate important variables that affect the insertion process, such as the regional variation of TA dinucleotide densities in the genome, and then adapt the method to find higher order combinations of mutations (CCIs) with improved stringency. The method is easily scalable and currently requires only a few minutes of CPU time. We use this model to analyze two insertional mutagenesis data sets, one from a transposon-based screen for gastrointestinal tract (GI) cancer genes and the other from the collection of retrovirus insertion data sets mainly in hematopoietic and brain tumors cataloged in the RTCGD. Finally, to gauge the superiority of the PRIM we compare the CIS gene sets from previous methods to the PRIM CISs by analyzing their overlap with human tumor data. Using three different human tumor data sets, we show that the PRIM method selects CISs with greater biological relevance compared to previous methods.

METHODS

Data preparation steps

The transposon insertion data set we analyzed was generated in an SB transposon-based forward genetic screen for mutations that cause GI tract cancer in mice (10). This study analyzed 16 690 non-redundant transposon insertions found in 135 GI tract tumors. We eliminated 100 insertions because of clonal duplications or their appearance in a common insertion site from a control data set of mouse tail insertions, leaving a set of 16 590 insertions. We also eliminated the 733 insertions on the sex chromosomes in order to simplify our presentation of the method, although the method can easily be adjusted for gender to include these insertions. The final data set contained 15 857 non-redundant transposon insertions in the autosomal chromosomes (see Supplementary Data set S1).

The mouse genome was divided into equally sized windows of fixed width and then the number of insertions counted within each window. The insertion counts can be recorded for various window sizes. The analysis in this article will use window sizes in 10 kb increments in the range from 10 to 150 kb with particular focus on the sizes 20, 50, 70 and 100 kb. These window sizes were chosen in order to compare our proposed CIS detection methods with those described previously (9,10). An analysis of each individual window size indicates that no single window identifies >50% of the union list of CISs (Supplementary Figure S1). In addition, although the majority of CISs in the union list are identified by multiple window sizes, 30% of CISs in the union list are only found with a single window size (Supplementary Figure S2). Because of these findings we believe it is

Table 1. Number of windows containing the indicated number of transposon insertions and the subset identified as statistically significant CISs[a] in the GI tumor data set

Window size (kb)	Total number of windows[b]	Number of windows with indicated number of insertions (number of statistically significant CIS windows)									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
20	120 774	106 582 **(0)**	12 782 **(0)**	1235 **(0)**	145 **(0)**	21 **(0)**	2 **(2)**	0 **(–)**	1 **(1)**	2 **(2)**	0 **(–)**
50	48 313	35 910 **(0)**	9773 **(0)**	2080 **(0)**	400 **(0)**	98 **(0)**	33 **(8)**	11 **(11)**	3 **(3)**	1 **(1)**	0 **(–)**
70	34 516	23 034 **(0)**	8319 **(0)**	2364 **(0)**	565 **(0)**	156 **(0)**	45 **(2)**	20 **(18)**	8 **(8)**	0 **(–)**	1 **(1)**
100	24 161	13 703 **(0)**	6859 **(0)**	2458 **(0)**	764 **(0)**	247 **(0)**	75 **(0)**	28 **(7)**	13 **(11)**	9 **(9)**	1 **(1)**
		10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**28**	**43**
20		0 **(–)**	1 **(1)**	0 **(–)**	1 **(1)**	1 **(1)**	1 **(1)**	0 **(–)**	0 **(–)**	0 **(–)**	0 **(–)**
50		0 **(–)**	0 **(–)**	0 **(–)**	1 **(1)**	0 **(–)**	1 **(1)**	1 **(1)**	0 **(–)**	1 **(1)**	0 **(–)**
70		1 **(1)**	0 **(–)**	0 **(–)**	0 **(–)**	0 **(–)**	1 **(1)**	0 **(–)**	0 **(–)**	2 **(2)**	0 **(–)**
100		0 **(–)**	1 **(1)**	0 **(–)**	0 **(–)**	0 **(–)**	0 **(–)**	1 **(1)**	1 **(1)**	0 **(–)**	1 **(1)**

[a]Number of statistically significant CISs based on the PRIM are in parentheses.
[b]Total number of windows in genome based on window size.

important to perform the PRIM at multiple window sizes and generate a union list from these analyses. Table 1 summarizes the insertion count and statistically significant CIS count for 4 of 15 different window sizes.

The retroviral insertion data set we analyzed comes from the publicly available RTCGD (described at http://variation.osu.edu/rtcgd). In order to compare our method to the Gaussian kernel convolution method, we used the same set of insertions that were used in the study by de Ridder *et al.* (14), which was kindly provided by the authors. This data set was generated from over 20 different studies and consists of 5473 insertions from 1361 tumors. Most of the tumors were hematopoietic (81%) and brain (8%) tumors.

Description of the PRIM method for detecting CISs

The number of times that insertions appear within a defined interval follows a Poisson process. Support for this assumption is provided in Section 2.1 in Supplementary Data. The Poisson probability distribution function is $P(X = x) = e^{-\lambda}\lambda^x/x!$ where the parameter λ is the rate of insertion and x is the number of insertions residing within a given window. The methods in this section assume that all windows within a single model are of the same size. The rate of insertion can account for other important variables, such as the number of TA sites, using a Poisson regression. For individual regions R_i, $i = 1, 2, \ldots, n_w$, where n_w is the number of windows of size w, the Poisson regression calculates the expected rate of insertion λ_i for region R_i using information about the size of the region, the chromosome it resides on, the number of TA sites within the window and the number of potential recoverable insertions. This last variable, the number of potential recoverable insertions, depends upon the restriction enzymes used during linker-mediated PCR (LM-PCR). A transposon insertion in a TA dinucleotide will not be recoverable if the nearest restriction enzyme cut site is too close to or too distant from the TA dinucleotide. The details for determining the TA dinucleotides and potential recoverable insertions (PRIs) in each window are provided in Section 1 of Supplementary Data. Figure 1 demonstrates that as window size increases, the insertion rate by chromosome varies increasingly. These chromosomal differences are accounted for by the coefficient β_c. The effect of the number of potential recoverable insertions is estimated with β_1. The TA sites are accounted for with an offset such that $\hat{\lambda}$ is roughly the number of insertions divided by the number of TA sites. The PRIM is

$$\log(\lambda_i) = \mu + \beta_c \mathrm{I}(R_i \in \text{Chromosome } c) + \beta_1 \times \text{PRI count} + \log(\text{TA sites in } R_i)$$

where $c = 1,2,\ldots,19$. The resulting fit from this regression will provide the expected rate of insertion for a given chromosome, a given number of PRIs and a given number of TA sites within each window. The Poisson regression above can be extended to account for other important variables such as mouse gender, donor concatemer site, or, when analyzing retroviral insertion data, transcription start site.

The expected rates of insertion $\lambda_1, \lambda_2, \ldots \lambda_{nw}$ are compared to the actual number of insertions within each window. Using the probability density function for the Poisson distribution given above, we can calculate the probability of the actual number of insertions given the expected rate from the Poisson regression $P(X \geq x_i|\lambda_i)$. To control the genome-wide false discovery rate (FDR) at 0.05, the resulting vector of probabilities needs to be adjusted for multiple comparisons. The Bonferroni correction of $0.05/n_w$ will select any window where the probability of observing at least x_i insertions is less than the Bonferroni corrected probability. For 20 kb windows, this threshold is 4.17×10^{-7} and for 100 kb windows, this threshold is 2.08×10^{-6}. The Bonferroni correction to select the threshold for significance is the most conservative choice to control the genome-wide FDR. If instead we employ a more liberal strategy to control the genome-wide FDR, such as the method of Benjamini and Hochberg (15,16), this will increase the detection of regions with a significant number of insertions.

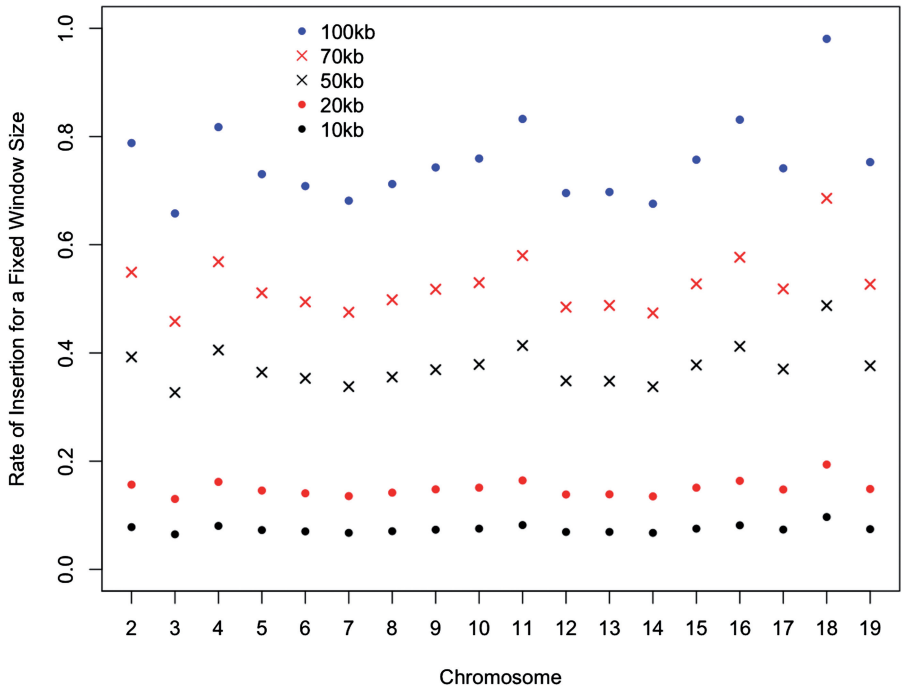

Figure 1. For various window sizes, a plot of the average rate of insertion for each mouse chromosome using the 15 857 insertions from the Starr *et al.* (10) data set. Conceptually, the rate parameter reflects the number of insertions per window, adjusting for the TA count. Chromosome 1 was dropped from the plot because for many mice this was where the donor transposon concatamer resided. All insertions that appeared on the same chromosome as their donor concatamer were removed in order to eliminate local-hopping artifacts. The local-hopping phenomenon is explained in more detail in Starr *et al.* (10).

Description of the PRIM method for detecting CCIs

Using the same definition as de Ridder *et al.* (14), a co-occurrence is a unique combination of insertions within a single tumor. To determine the observed rate of common co-occurrence, one counts the number of times a unique combination of insertions is observed in a set of tissue samples. If insertions occur within the same pair of windows within two or more tissue samples, this is an observed co-occurrence.

The PRIM method introduced above can easily be extended to model the expected rate of co-occurrence for any two pairs of windows. Recall that, in the previous section, we defined the rate λ_i as the rate of insertion into a chromosome for a given number of TA dinucleotides and given window size. If we observe insertions along one axis, representing the first insertion in the pair, with a rate of λ_i, and then observe insertions along another axis, representing the second insertion in the pair, with a rate of λ_j, then the expected rate of co-occurrence is approximately $\lambda_i\lambda_j$. The derivation of this rate is given in Section 2.2 of Supplementary Data.

With an expected rate of co-occurrence, we can determine the probability of observing the actual number of co-occurrences for a given pair.

$$P(X \geq x_{ij}|\lambda_i\lambda_j) = \sum_{X=x_{ij}}^{\infty} e^{-\lambda_i\lambda_j}(\lambda_i\lambda_j)^X/X!$$

This probability is only calculated for window pairs where there is a co-occurrence in at least two tumors. This means that the number of tests will vary depending on the window size and the insertion data set. Suppose there are p_w total window pairs in the data set of interest. Using a Bonferroni correction of $0.05/p_w$, we select any window pair where the probability of observing at least x_i co-occurrences is less than the Bonferroni corrected probability. If instead we employ a more liberal strategy to control the genome-wide FDR, such as the method of Benjamini and Hochberg, we may observe more window pairs with significant CCIs. Therefore, different strategies for FDR calculation will yield somewhat different results.

Description of the 2DGKC model for detecting CCIs

To date, there has been one published method to detect CCIs (14). The method fits a two-dimensional Gaussian Kernel Convolution (2DGKC) to estimate a bivariate density function. This obtains a smooth estimate of the number of insertion co-occurrences for a given kernel width. The smoothed estimate is a continuous function of count data. To determine statistical significance of a CCI, the approach performs multiple permutations at multiple kernel widths to generate a null distribution. At each permutation, a new density function is constructed and the calculated peak heights are summarized. The primary advantage of this method is the agnostic and non-parametric approach.

It was not possible, however, to use the 2DGKC method as published on the larger data set generated in our SB transposon screen because the computational requirements exceeded the capacity of our supercomputers. To overcome this limitation, we made two primary alterations to the 2DGKC approach. First, in addition to using the original permutation approach provided by the method's authors, we also introduce a modified permutation approach that more closely resembles the spirit of the permutation strategy suggested in de Ridder *et al.* (14). The modifications of this permutation strategy are more fully explained in the Section 5 of Supplementary Data. Second, in the original article, a co-occurring insertion was called significant if the actual peak height exceeded a threshold determined by a quantile of the peak heights averaged over 1000 permutations. In this article, we will more conservatively call a CCI significant if the actual peak height exceeds the maximum out of all peak heights generated in 1000 permutations.

We modified the MATLAB code that calculates CCIs using 2GKDC (kindly provided by de Ridder *et al.*) to improve its computational efficiency, and then used it to determine CCIs in the GI tumor data. To compute a null distribution, 1000 random runs were performed for each of the eight kernel widths used in their original study. These runs took ~8000 CPU hours on a Sun Fire X4600 Linux cluster which has 192 cores and 768 GB of memory. Using a modest level of parallelism (~20 CPUs), this computation can be completed in a couple of weeks on the machine just described. This is approximately the level of parallelism we used.

The data set of 16 590 insertions and 135 tumors translates into ~1.6 million two-dimensional points, one for each possible pair of insertions in a tumor. Optimization techniques are used to find the peaks in the density generated by these points as described in de Ridder *et al.* (14). Our modifications to the code preserved this approach, but reduced the time and/or memory requirements of (i) the use of the MATLAB optimization package, (ii) the grouping of points into bins and the identification of points in neighboring bins and (iii) the filtering of redundant peaks. Section 4 of Supplementary Data provides further details about these modifications. While the original code could not calculate peaks on our data set in a finite amount of time, due to memory allocation problems, modifications to the code fixed this issue. Further simple modifications reduced computation time by 90%. More extensive improvements resulted in an additional speed up of roughly two orders of magnitude. Likewise, for the creation of random data for generating the null distribution, we followed the original approach, but significantly improved computational efficiency. The modified code was tested on the RTCGD data set to ensure that it yielded results that were the same as the original code. Despite improvements in computational efficiency, simulations to estimate the empirical FDR for the 2DGKC methods were not feasible. The computation time for this approach is compared to PRIM for the GI tumor data in Supplementary Table S1. This table indicates that PRIM requires ~0.01% of the CPU time on the same machine that the 2DGKC method with permutations requires.

Simulation methods

In order to obtain an empirical estimate of the FDR or specificity of the PRIM methods, we conducted simulation exercises. Monte–Carlo simulations were performed to generate null distributions, where insertions are scattered completely at random. To create these simulation data sets, a set of insertions is first randomly allocated to a set of tumors. The number of insertions per tumor will vary empirically. Then the insertions are randomly linked with 1 of over 150 million TA dinucleotides across the genome. After creating the set of insertions, the insertion sites and co-occurrences are counted and PRIM tests for significance. All simulations assume window sizes of 100 kb. The process is repeated 1000 times to obtain an empirical estimate of the FDR.

Simulations were initially run for nine distinct scenarios. The samples had a size of 7500, 15 000 or 30 000 insertions. These insertions were randomly assigned to either 50, 100 or 150 tumors. In addition, a tenth simulation scenario mimicked the insertion set from Starr *et al.* (10). Thus, for each random sample, 16 000 insertions were randomly allocated to 135 tumors. The entire set of simulations can be generated and analyzed in <1 week using three Linux servers with 2 GHz processors.

RESULTS

The PRIM detects CISs in GI tumors

To compare the PRIM method to Monte–Carlo simulations previously used to identify CISs, we re-analyzed a data set of SB transposon insertions in GI tract tumors in mice (10). Analysis was done using genomic window sizes from 10 to 150 kb in steps of 10 kb. The number of genomic windows and the number of transposon insertions within those windows is listed in Table 1 for 4 of the 15 window sizes analyzed. The subset of these windows that were statistically significant CISs (adjusted $P < 0.05$) is indicated in parenthesis. From the table, it is apparent that any window harboring eight or more insertions, independent of window size, is a CIS.

In addition to window size, the analysis also factored in the number of TA dinucleotides present in each genomic window based on the most recent mouse genome build (NCBIM37). This was done because the SB transposon only inserts into TA dinucleotides (17). We found that the number of TA dinucleotides in a window can vary substantially. For example, in 100 kb windows the minimum TA dinucleotide count within a window was 31 and the maximum was 9376 sites. This demonstrates that the expected rate of insertion can be 300 times as large for one window compared to another. Accounting for TA dinucleotides in PRIM led to changes in the composition of the CIS lists by up to 18 different regions, while accounting for PRIs only changed the composition by 2 regions (see Supplementary Tables S2 and S3). By identifying the union of all CISs defined by PRIM, which accounts for both TA density and number of PRIs,

using the 15 different window sizes there were either 28 or 88 CISs depending upon whether the Bonferroni or the Benjamini and Hochberg adjustment for multiple comparisons was used (Supplementary Tables S2 and S3) (15,16). In the original study, using the same data set, there were 61 CISs defined by Monte–Carlo simulations using a threshold E-value <1 (Supplementary Table S4). We analyzed the overlap between CISs found by Monte–Carlo simulations versus PRIM with the Benjamini and Hochberg adjustment and found that 59 of the 61 CISs identified by Monte–Carlo simulations were also identified by PRIM.

The PRIM method discovers more relevant cancer genes. It is difficult to compare the benefits of using different methods for detecting CISs because the role of the majority of genes in the genome is still not well characterized (18). To gauge which method may be producing more relevant results, we used three tests to compare the biological relevance of the different CIS sets to data sets analyzing human tumors. First, we tested the gene sets identified by the CIS lists using the Globaltest methods proposed in Goeman *et al*. (19) to see if there was differential expression based on cDNA microarray data published in Ki *et al*. (20). Using matched primary human colon tumors and normal mucosa for 23 of the patients with expression data, we can examine differential expression on pre-specified sets. Employing the 'self-contained hypothesis' for testing, the Q statistic used in the Globaltest methods determines if there is differential expression in the gene set under inquiry. The greater the differential expression in the set, the larger the Q statistic is. All of the sets of CIS genes queried were statistically significant. The set of 59 CISs determined by PRIM that overlapped with the previously published list had the largest Q statistic, suggesting the largest effect size and the most differential expression. The second method we used to compare the biological relevance of the different gene sets was to calculate the overlap between the CIS lists and the cancer gene census, which is a census of known, bona fide cancer genes (21). In this comparison, all sets of CIS genes were statistically significant and the PRIM CIS lists had the highest percentage of overlap. Finally, we also explored the overlap between the CIS lists and a set of genes identified as CRC cancer genes based on re-sequencing of ~18 000 genes in 11 human CRCs (22). In that study, the authors identified 140 genes that were mutated in human CRC and likely contributed to the tumor growth. The CIS lists found by PRIM showed a statistically significant overlap with these genes ($P = 0.012$) while the CIS list generated using Monte–Carlo simulations (10) did not ($P = 0.132$) (see Section 3 of Supplementary Data). In addition, the CIS found by PRIM that are not found in the Monte–Carlo simulations also show a statistically significant overlap with the CRC cancer genes ($P = 0.003$). These three lines of evidence suggest that the PRIM method may identify more relevant cancer genes than previous methods based on Monte–Carlo simulations.

The PRIM detects CCIs in GI tumors

The CIS analysis presented above identifies single mutations that contribute to tumor formation. We now expand this analysis to identify pairs of mutations that cooperate in tumorigenesis. The PRIM for CCIs detects pairs of regions containing insertions at a rate higher than expected by chance. Table 2 shows the number of these pairs of insertions appearing together in two or more tumors in this data set for each of four window sizes. For example, pairs of 20 kb windows both containing a transposon insertion were found 605 times in two tumors, while one pair was found in three tumors. The majority of co-occurrences are found in two tumors and, as would be expected, the number increases as window size increases. Regardless of window size, a small number of co-occurrences are found in three or more tumors. The largest number of co-occurrences observed is 9137 pairs when the window size is 100 kb. To filter out CCIs that occur by random chance, we use a Bonferroni correction of $0.05/9137 = 5.5 \times 10^{-6}$ and select any 100 kb window where the probability of the observed CCI is less than the threshold. This identified three significant CCI pairs with a greater number of tumors than expected by chance (Table 3). The genome-wide FDR correction of Benjamini and Hochberg detects the same three CCIs.

In order to compare our method of CCI detection to the method developed by de Ridder *et al*. (14), we fit a two-dimensional Gaussian Kernel Convolution (2DGKC) to the set of GI tumor insertions to derive a list of CCI candidates based on this approach. At each bandwidth, the threshold for significance was chosen as the maximum peak height found in 1000 permutations of the insertion set. This threshold choice is slightly more conservative than the threshold advocated in the original article (14), where they used a quantile of peak heights that were averaged over 1000 permutations. For each of the eight bandwidths employed, an increasing number of statistically significant CCIs were found as the bandwidth size increased (Supplementary Table S5). For the smallest bandwidth of size 10 kb, the method detects 56 CCIs. For the largest bandwidth of size 500 kb, the method detects 1176 CCIs. Since the 2DGKC method smoothes over count data to construct a continuous function, some of the significant CCIs in the resulting list do not contain any co-occurring insertions. These instances are shown in Supplementary Table S5 where the tumor count is zero or one.

Table 2. Number of CCIs in a given genomic window in 2, 3, 4, 5 or 7 tumors from the GI tumor data set

Window size (kb)	Number of tumors				
	2	3	4	5	7
20	605	1	0	0	0
50	2972	26	2	0	0
70	5235	55	1	1	0
100	9009	121	6	1	1

Table 3. Three CCIs occurring in GI tumor data set

Window size[b] (kb)	Locus A of CCI[a]			Locus B of CCI[a]			Number of tumors
	Chr	Start address[c]	Gene name[d]	Chr	Start address[c]	Gene name[d]	
70	10	122 140 001	Ppm1h	15	42 970 001	Rspo2	4
100	5	148 300 001	Pan3	11	86 500 001	Cltc	4
100	18	34 300 001	Apc	18	34 400 001	Apc	7

[a]Locus A and Locus B are the pair of loci composing the CCI.
[b]When multiple window sizes find the same CCI, the largest window size is reported.
[c]Physical address of start of CCI based on NCBIM37 genome build.
[d]Candidate gene in locus.
Chr = Chromosome.

Table 4. Number of co-occurring insertions and the subset identified as statistically significant CCIs[a] in the RTCGD

Window size (kb)	Number of tumors with a co-occurring pair of insertions							
	2	3	4	5	6	7	8	9
20	49 **(2)**	4 **(4)**	1 **(1)**	1 **(1)**	0 **(–)**	0 **(–)**	0 **(–)**	0 **(–)**
50	64 **(0)**	10 **(10)**	2 **(2)**	1 **(1)**	0 **(–)**	0 **(–)**	1 **(1)**	0 **(–)**
70	82 **(0)**	7 **(7)**	2 **(2)**	2 **(2)**	1 **(1)**	0 **(–)**	0 **(–)**	1 **(1)**
100	98 **(0)**	10 **(10)**	4 **(4)**	3 **(3)**	0 **(–)**	0 **(–)**	0 **(–)**	1 **(1)**

[a]Number of statistically significant CISs based on the PRIM are in parenthesis.

The originally published permutation strategy for the 2DGKC method samples random pairs of insertions from locations in the genome. Alternatively, we tested a new permutation strategy where we preserved the number and location of the original insertions, and the number and size of the original tumors, but merely permuted the tumor labels on the insertions. This permutation strategy preserves the structure of the original tumor data, but eliminates the relationship between pairs of insertions (See Section 4 of Supplementary Data). Interestingly, when using our modified version of the permutation strategy, instead of the original, there were no CCIs detected at any bandwidth. This demonstrates that the choice of permutation strategy greatly affects CCI detection.

The list of three statistically significant CCIs found via PRIM was cross-referenced with ranked lists of 2D peaks at each bandwidth based on the original permutation strategy. The CCI in Table 3, where both pairs occur within the *Apc* gene, is the most significant CCI in the list of 10 kb bandwidth peaks. The CCI between *Rspo2* and *Ppm1h* lies in the bottom 25% of significant CCIs found using a bandwidth of either 17.5 or 30.6 kb. The CCI between *Pan3*/*Flt1* and *Cltc* falls in the top 10% of significant CCIs found by 286 or 500 kb bandwidths. The differences between the methods under investigation reflect (i) the differing choice in assuming a count variable versus a continuous variable to model insertions and (ii) the differing choice of null distribution assumptions.

The PRIM detects CCIs in RTCGD tumors

Next, we applied our method to a data set of retroviral insertions in mouse tumors. Using the 2DGKC method, de Ridder *et al.* (14) reported 86 statistically significant CCIs. Analysis of the same data set using PRIM detected only 20 statistically significant CCIs out of 116 total co-occurrences using window sizes of 20, 50, 70 and 100 kb (Table 4 and Supplementary Table S6), indicating PRIM is four times more conservative than the 2DGKC method. Nineteen of the 20 CCIs detected by PRIM were also detected by the 2DGKC method. Based on the ranking of the 86 CCIs detected by 2DGKC, PRIM detects only highly ranked CCIs as all 19 CCIs were ranked in the top 36% of the original 86 reported CCIs ($P < 0.001$) (Supplementary Table S6).

FDR of PRIM using simulated data

To obtain an empirical estimate of the FDR or specificity of PRIM, we generated data sets of random insertions throughout the mouse genome. By definition, any significant CCIs in these simulated data sets are false discoveries. We analyzed 1000 simulations within each of 10 different combinations of tumor and insertion counts. We included the combination of 135 tumors and 16 000 insertions because it is the equivalent size of the GI tract data set analyzed in 'The PRIM detects CCIs in GI tumors' section. Since PRIM is calibrated to control the FDR at 0.05, corresponding to 50 simulations out of 1000 with a significant CCI, this is the ideal rate in our simulations. When using the Benjamini and Hochberg method for multiple comparisons adjustment, the FDRs were <0.05 (Table 5).

Any statistical technique will inevitably have a certain level of false discoveries. These false discoveries should occur, however, when the characteristics of the CCIs in the random data match most closely with genuine CCIs in the real data. Specifically, the more tumors in which the

Table 5. Empirical FDR of PRIM using various sized simulated data sets[a]

Number of tumors	Number of insertions			
	7500	15 000	16 000	30 000
50	0.036	0.011	ND[b]	0
135	ND[b]	ND[b]	0.002	ND[b]
150	0.016	0.002	ND[b]	0.001
300	0.002	0	ND[b]	0

[a]FDR is based on 1000 simulations and is calculated as (number of simulations that produced a CCI)/(number of total simulations). In all individual simulations that identified a CCI, only one CCI was found, except in one of the simulations with 50 tumors and 7500 insertions, where one of the simulations yielded two CCIs.
[b]ND indicates simulations were not done.

Table 6. Number of simulations with pairs of insertions in the indicated number of tumors[a]

Number of tumors	Number of insertions			
	7500	15 000	16 000	30 000
50	681 (2) 319 (3)	944 (3) 56 (4)	ND[b]	935 (4) 65 (5)
135	ND[b]	ND[b]	130 (2) 868 (3) 2 (4)	ND[b]
150	984 (2) 16 (3)	374 (2) 624 (3) 2 (4)	ND[b]	380 (3) 168 (4) 2 (5)
300	998 (2) 2 (3)	899 (2) 101 (3)	ND[b]	992 (3) 8 (4)

[a]For example, in 1000 simulations using 50 tumors and 7500 insertions there are 681 simulations with a pair of insertions occurring in 2 tumors and 319 simulations with a pair of insertions occurring in 3 tumors.
[b]ND indicates simulations were not done.

pair of insertions comprising the CCI occur together, the more likely it is that the CCI is not due to random chance. Thus, we investigated whether the false discoveries generated by our approach tended to be those CCIs in the random data that occur in the largest number of tumors. To this end, Table 6 shows the distribution of tumor counts with insertion pairs from each simulation scenario. The tumor count is the maximum number of tumors in which a pair occurs. For instance, for the scenario with 50 tumors and 7500 insertions, 681 of the 1000 simulations have pairs in at most two tumors, while 319 of the simulations have pairs in at most three tumors.

In all scenarios, the simulations with significant CCIs come from simulations that have the maximum tumor count for their scenario. Indeed, in 4 of the 10 scenarios, the false discovery simulations are exactly those simulations with the maximum number of CCI tumor pairs. To illustrate, in the simulation scenario where 16 000 insertions are distributed among 135 tumors, the maximum CCI count is 4 tumors with the same CCI pair, which occurs in only 2 of the 1000 simulations. These same two simulations are also the only two simulations that result in a significant CCI call for this scenario. Meanwhile, a maximum CCI count of three tumors occurs in 868 of the 1000 simulations and two tumors in the remaining 130 simulations. Therefore, for this particular simulation scenario, an FDR of 2 out of 1000 is reasonable.

When comparing the simulation scenario to the actual data this scenario mimics, i.e. the GI tumor insertions, the simulations support the notion that CCI pairs within 100 kb windows appearing in four tumors or less are to be expected by chance. PRIM detects two significant CCIs in 100 kb windows, one appearing in five tumors (*Apc*|*Apc*) and one appearing in four tumors (*Pan3*|*Cltc*). The remaining five CCIs detected in four tumors in the 100 kb window analysis have a more typical TA count compared to the significant *Pan3*|*Cltc* pair and did not cross the threshold for significance.

Although the simulation FDRs are all <0.05, this does not imply that our methods are too conservative or too specific. On the contrary, our simulations show that rare events are detected when they occur, just as Poisson models are designed to do. Table 6 indicates that these rare events of occurrence appear under the null distribution at a rate <0.05 and thus the simulation FDRs are also <0.05. This suggests that the proposed methods appropriately control the genome-wide FDR and detect the correct number of CCIs.

DISCUSSION

We have developed a new method to identify CISs and CCIs that play a role in tumorigenesis. This approach uses a two-dimensional extension of rate parameter estimates found with a Poisson regression model. The assumptions of PRIM were verified with simple Monte–Carlo simulations.

Several of the genes identified by PRIM, but not by the Monte–Carlo method, are known to be involved in the pathogenesis of human colorectal cancer (CRC). For example, *PIK3CA*, the catalytic unit of phosphatidylinositol-3 kinase (PI3K), is mutated in 32% of human colon cancers (23), and up to 40% of CRCs have mutations in PI3K pathway genes (24). Mutant *PIK3CA* promotes cell growth and tumor invasion and enhances metastatic CRC resistance to treatment by monoclonal antibodies targeting *EGFR* (25,26). PRIM with the TA site offset, but not the null model, identified the ephrin receptor, *EPHB2*, which is initially upregulated in early colon lesions but is subsequently downregulated as the tumor progresses and this silencing correlates with poor patient survival (27). The tumor suppressor function of EPHB2 may be responsible for hereditary CRC due to a germline mutation in rare cases (28). Several of the CISs identified by PRIM exhibit altered expression in colorectal cancer, such as *MUC5AC* (29). These findings support the validity of the PRIM method to find candidate cancer genes using transposon insertional data sets, where these genes may be missed using the previously employed Monte–Carlo method.

We then extended PRIM to allow detection of CCIs. The PRIM method to detect CCIs was tested in simulations of the null hypothesis, a set of gastrointestinal tumors, and a set of tumors from the RTCGD. We demonstrated in simulations that the PRIM method for CCI detection properly controls the FDR. The empirical FDRs in 10 different simulation scenarios were always <0.05. The simulated data also showed the number of CCIs that are expected by chance. Given that PRIM did not select pairs in the GI tumor set with a smaller number of counts than observed in simulations, this shows that the model appropriately controls false discoveries. On the other hand, PRIM did select CCIs in the GI tumors with a larger number of counts than observed in simulations, showing that the model is not overly conservative.

After verification with simulations, we then used our methods to determine CCIs in a set of GI tumors from mice and found three statistically significant CCI regions. The basic hypothesis for doing a CCI analysis is that cooperating mutations can be identified. An analysis of the function of the CCI pairs identified supports this hypothesis.

The first CCI identified cooperating mutations within the *Apc* gene. The biology of *Apc* mutations has been extensively studied in colon cancer and the current hypothesis is that an initial truncating mutation resulting in a hypomorphic protein product is normally followed by loss of heterozygosity of the remaining wild-type allele (30). This hypothesis fits well with the *Apc-Apc* CCI identified in the set of intestinal tumors analyzed in this article. The seven tumors that constitute this CCI show multiple mutations in the first intron and 3′ upstream region of *Apc* accompanied by paired mutations in downstream introns (Figure 2A). It is possible that the first mutation creates a null product because it occurs in the first intron or promoter region of *Apc*, while the second mutation creates a truncated product in the second allele.

The second CCI identified *Rspo2* and *Ppm1h* as interacting mutations. This could be explained by the hypothesis that overexpression of *Rspo2* and inactivation of *Ppm1h* cooperate in CRC progression by fulfilling two of the functions associated with mutant *Apc*, namely uncontrolled proliferation and chromosomal instability. In addition to activation of Wnt signaling, mutations in the C-terminus of *Apc* contribute to chromosomal instability (31). This second function may explain why *Apc* mutations are found more frequently in CRC than other genes capable of activating Wnt signaling. The four tumors constituting the *Rspo2-Ppm1h* CCI do not have any identified transposon insertions in *Apc*, so the *Rspo2* and *Ppm1h* mutations could be providing the phenotypes usually caused by mutant *Apc*. The transposon insertions in *Rspo2* in the four tumors are likely causing overexpression because the insertions are all located

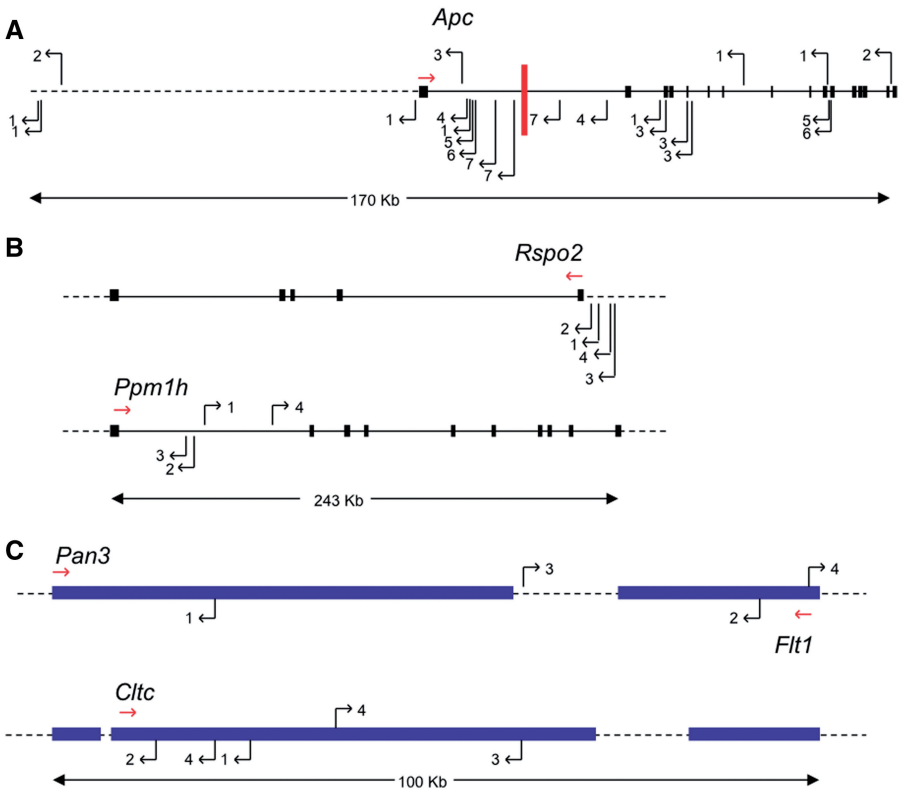

Figure 2. Location of transposon insertions in CCIs. (**A**) Seven tumors had insertions in both the 3′ and 5′ regions of Apc. (**B**) Four tumors had insertions in both the upstream promoter of Rspo2 and in intron 1 of Ppmh1. All four insertions in the Rspo2 promoter inserted with the transposon viral promoter in the same orientation as the gene. (**C**) Four tumors had insertions in Cltc, two of which had insertions in Flt1 and the other two in Pan3. Insertions are depicted by a bent arrow, which points in the direction of the transposon promoter. Insertion numbers indicate tumors. Red arrow indicates direction of transcription. Solid black lines depict introns while dashed black lines depict intergenic DNA. Black boxes depict exons (A and B) while blue boxes depict genes (C). Arrow on bottom indicates length of DNA.

immediately upstream of *Rspo2* and the viral promoter within the transposon is in the correct orientation to cause overexpression (Figure 2B). These insertions are probably causing aberrant activation of Wnt signaling because *Rspo2* normally functions as a secreted activator of the Wnt signaling pathway that is important for limb, lung and craniofacial development (32–35). Furthermore, *Rspo1*, a close homolog of *Rspo2* causes hyperproliferation in intestinal crypt cells along with an increase in β-catenin levels when the human protein is overexpressed in mice (36). The transposon insertions in *Ppm1h*, on the other hand, are likely causing disruption of this gene because they are spread throughout the gene and the direction of the viral promoter is not consistent (Figure 2B). Inactivation of *Ppm1h* could be cooperating with overexpression of *Rspo2* by interfering with p53 transcription leading to chromosomal instability. *Ppm1h* is a protein phosphatase that can dephosphorylate and potentially inactivate *CSE1L* (37). *CSE1L* was recently shown to be associated with chromatin and to regulate transcription of p53 target genes (38). Furthermore, *CSE1L* intracellular localization is controlled by phosphorylation and *CSE1L* will accumulate in the nucleus when phosphorylation is blocked (39). Based on these observations we predict that overexpression of *Rspo2* and inactivation of *Ppm1h* cooperate in the etiology of CRC.

The third CCI, designated *Pan3-Cltc*, contains three affected genes. All four tumors had transposon insertions in Clatherin heavy chain (*Cltc*) while two of the tumors had insertions in FMS-like tyrosine kinase 1 (*Flt1*, *alias Vegfr1*), one had an insertion in the neighboring gene PAN3 polyA specific ribonuclease subunit homolog (*Saccharomyces cerevisiae*) (*Pan3*), and one had an insertion in the intergenic region between *Pan3* and *Flt1* (Figure 2C). From the insertion pattern, we predict that *Cltc* is inactivated in all four tumors, while it is difficult to predict the effect of the insertions on *Pan3* and *Flt1*. It is possible that the *Flt1* mutations create a truncated product, as the insertions are located toward the 5′ end of the *Flt1* gene. This might result in a protein product similar to the shortened, soluble isoform *sFlt1*. Although there is evidence that delivery of *sFlt1* using gene therapy can block tumor development in mouse models (40), increased levels of *sFlt1* are found in the sera of colorectal and breast cancer patients (41) and elevated *sFlt1* levels in renal cancer are associated with a poorer outcome (42). These findings suggest there may be an oncogenic component to *sFlt1*. Inactivation of *Cltc* might be contributing to tumor development due to increased *Egfr* signaling. Activated *Egfr* is normally targeted for destruction after ubiquitination and subsequent transport from the plasma membrane to lysosomes. *Cltc* controls *Egfr* signaling by acting as a chaperone transporting activated *Egfr* to the lysosome (43). Loss of *Cltc* may result in prolonged *Egfr* signaling leading to uncontrolled proliferation, which could cooperate with dysregulation of *Vegf* signaling due to the mutations in *Flt1*.

The three significant CCI regions found by PRIM potentially explain part of the tumorigenesis stages of as many as 11 tumors in the data set, out of 135 tumors total. The 88 CIS regions found by the Poisson model involve insertions from 117 of the tumors. Most, though not all, of the tumors in the data set may in part be explained by one or two disruptions due to SB insertions. It is likely that a complete picture of tumorigenesis will require a model with more than two hits.

Comparing PRIM to the existing 2DGKC method, we found that PRIM is far more discerning. The 2DGKC method found 1176 CCIs in the GI tumors. In the RTCGD insertion set, <25% of the CCI regions detected by de Ridder *et al*. (14) were found to be significant by our model. Modifying the permutation strategy that generates peaks under a null distribution greatly reduces the number of CCIs detected (see the Section 4 of Supplementary Data for more description of the permutation strategies). This suggests that inference under the 2DGKC method could be more similar to PRIM when using improved permutation strategies.

The PRIM framework provides for more flexibility in the estimation of transposon insertion rates. This means that as the process of transposon-based screens are better understood with time, we will be able to easily include new variables that affect the rate of insertion. We are currently expanding the methods proposed in this article to accommodate mouse gender and donor concatemer site in the model and therefore we will be able to analyze insertions on sex chromosomes and account for the local-hopping phenomenon without bias. The new methods for CCI detection are also far faster to compute than previous methods. The more efficient computations allow us to verify our approach with simulations, whereas the previously published approaches do not. The code in R to calculate the rate of insertion and co-occurrence and identify CISs and CCIs is available upon request (http://www.r-project.org). The ease of computation also provides future opportunities to expand our approach to higher order combinations of insertions beyond a two-hit model.

In conclusion, we have presented a new method for determining CISs and CCIs from data sets of transposon or proviral insertions in forward genetic screens for cancer genes. The new method, termed PRIM, is able to identify the biologically relevant mutations in these screens and can be tailored to screen-specific behaviors such as the requirement of TA dinucleotides for SB transposons or the preference of proviruses to insert into TSSs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S7, Supplementary Figures S1–S5, Supplementary Methods, Supplementary Data set S1 and Supplementary Reference [44].

ACKNOWLEDGEMENTS

The authors would like to thank Jeroen de Ridder for helpfully providing the MATLAB code and documentation that execute the original 2DGKC methods. They also thank the Minnesota Supercomputing Institute for

computational infrastructure and systems administration support and the Masonic Cancer Center Biostatistics and Bioinformatics Shared Resource.

FUNDING

National Institute of Health (grant number R01-CA113636); Post-doctoral Fellowship from the American Cancer Society (grant number PF-06–292-01-MGO to T.K.S.); NIH Pathway to Independence award (grant number 4 R00CA151672-03 to T.K.S.) and National Science Foundation (grant number III:0916439). Funding for open access charge: University of Minnesota.

Conflict of interest statement. None declared.

REFERENCES

1. Ashley,D.J. (1969) The two "hit" and multiple "hit" theories of carcinogenesis. *Br. J. Cancer*, **23**, 313–328.
2. Knudson,A.G. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci.*, **68**, 820–823.
3. Knudson,A.G. (2001) Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer Sci.*, **1**, 157–162.
4. Fearon,E.R. and Vogelstein,B. (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.
5. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
6. Collier,L.S., Carlson,C.M., Ravimohan,S., Dupuy,A.J. and Largaespada,D.A. (2005) Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*, **436**, 272–276.
7. Dupuy,A.J., Akagi,K., Largaespada,D.A., Copeland,N.G. and Jenkins,N.A. (2005) Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*, **436**, 221–226.
8. Dupuy,A.J., Rogers,L.M., Kim,J., Nannapaneni,K., Starr,T.K., Liu,P., Largaespada,D.A., Scheetz,T.E., Jenkins,N.A. and Copeland,N.G. (2009) A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res.*, **69**, 8150–8156.
9. Keng,V.W., Villanueva,A., Chiang,D.Y., Dupuy,A.J., Ryan,B.J., Matise,I., Silverstein,K.A., Sarver,A., Starr,T.K., Akagi,K. *et al.* (2009) A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat. Biotechnol.*, **27**, 264–274.
10. Starr,T.K., Allaei,R., Silverstein,K.A., Staggs,R.A., Bergemann,T.L., O'Sullivan,M.G., Matise,I., Dupuy,A.J., Collier,L.S., Powers,S. *et al.* (2009) A transposon-based genetic screen identifies genes altered in colorectal cancer. *Science*, **323**, 1747–1750.
11. Uren,A.G., Kool,J., Berns,A. and van Lohuizen,M. (2005) Retroviral insertional mutagenesis: past, present and future. *Oncogene*, **24**, 7656–7672.
12. Mikkers,H., Allen,J., Knipscheer,P., Romeijn,L., Hart,A., Vink,E. and Berns,A. (2002) High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat. Genet.*, **32**, 153–159.
13. de Ridder,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**, e166.
14. de Ridder,J., Kool,J., Uren,A., Bot,J., Wessels,L. and Reinders,M. (2007) Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. *Bioinformatics*, **23**, i133–i141.
15. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.
16. Ferriera,J.A. and Zwinderman,A.H. (2006) On the Benjamini-Hochberg Method. *Ann. Stat.*, **34**, 1827–1849.
17. Ivics,Z., Hackett,P.B., Plasterk,R.H. and Izsvák,Z. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501–510.
18. Huss,J.W. III, Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., Hogenesch,J.B. and Su,A.I. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **38**, D633–D639.
19. Goeman,J.J., van de Geer,S.A., de Kort,F. and van Houwelingen,H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
20. Ki,D.H., Jeung,H.C., Park,C.H., Kang,S.H., Lee,G.Y., Lee,W.S., Kim,N.K., Chung,H.C. and Rha,S.Y. (2007) Whole genome analysis for liver metastasis gene signatures in colorectal cancer. *Int. J. Cancer*, **121**, 2005–2012.
21. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
22. Wood,L.D., Parsons,D.W., Jones,S., Lin,J., Sjoblom,T., Leary,R.J., Shen,D., Boca,S.M., Barber,T., Ptak,J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
23. Samuels,Y., Wang,Z., Bardelli,A., Silliman,N., Ptak,J., Szabo,S., Yan,H., Gazdar,A., Powell,S.M., Riggins,G.J. *et al.* (2004) High frequency of mutations of the PIK3CA gene in human cancers. *Science*, **304**, 554.
24. Parsons,D.W., Wang,T.L., Samuels,Y., Bardelli,A., Cummins,J.M., DeLong,L., Silliman,N., Ptak,J., Szabo,S., Willson,J.K. *et al.* (2005) Colorectal cancer: mutations in a signalling pathway. *Nature*, **436**, 792.
25. Samuels,Y., Diaz,L.A. Jr, Schmidt-Kittler,O., Cummins,J.M., Delong,L., Cheong,I., Rago,C., Huso,D.L., Lengauer,C., Kinzler,K.W. *et al.* (2005) Mutant PIK3CA promotes cells growth and invasion of human cancer cells. *Cancer Cell*, **7**, 561–573.
26. Sartore-Bianchi,A., Martini,M., Molinari,F., Veronese,S., Nichelatti,M., Artale,S., Di Nicolantonio,F., Saletti,P., De Dosso,S., Mazzucchelli,L. *et al.* (2009) PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer Res.*, **69**, 1851–1857.
27. Clevers,H. and Batlle,E. (2006) EphB/EphrinB receptors and Wnt signaling in colorectal cancer. *Cancer Res.*, **66**, 2–5.
28. Zogopoulos,G., Jorgensen,C., Bacani,J., Montpetit,A., Lepage,P., Ferretti,V., Chad,L., Selvarajah,S., Zanke,B., Hudson,T.J. *et al.* (2008) Germline EPHB2 receptor variants in familial colorectal cancer. *PLoS One*, **3**, e2885.
29. Forgue-Lafitte,M.E., Fabiani,B., Levy,P.P., Maurin,N., Fléjou,J.F. and Bara,J. (2007) Abnormal expression of M1/MUC5AC mucin in distal colon of patients with diverticulitis, ulcerative colitis and cancer. *Int. J. Cancer*, **121**, 1543–1549.
30. Nieuwenhuis,M.H., Mathus-Vliegen,L.M., Slors,F.J., Griffioen,G., Nagengast,F.M., Schouten,W.R., Kleibeuker,J.H. and Vasen,H.F. (2007) Genotype-phenotype correlations as a guide in the management of familial adenomatous polyposis. *Clin. Gastroenterol. Hepatol.*, **5**, 374–378.
31. Fodde,R. (2002) The APC gene in colorectal cancer. *Eur. J. Cancer*, **38**, 867–871.
32. Bell,S.M., Schreiner,C.M., Wert,S.E., Mucenski,M.L., Scott,W.J. and Whitsett,J.A. (2008) R-spondin 2 is required for normal laryngeal-tracheal, lung and limb morphogenesis. *Development*, **135**, 1049–1058.
33. Kazanskaya,O., Glinka,A., del Barco Barrantes,I., Stannek,P., Niehrs,C. and Wu,W. (2004) R-Spondin2 is a secreted activator of Wnt/beta-catenin signaling and is required for *Xenopus myogenesis*. *Dev. Cell*, **7**, 525–534.
34. Nam,J.S., Park,E., Turcotte,T.J., Palencia,S., Zhan,X., Lee,J., Yun,K., Funk,W.D. and Yoon,J.K. (2007) Mouse R-spondin2 is required for apical ectodermal ridge maintenance in the hindlimb. *Dev. Biol.*, **311**, 124–135.

35. Yamada,W., Nagao,K., Horikoshi,K., Fujikura,A., Ikeda,E., Inagaki,Y., Kakitani,M., Tomizuka,K., Miyazaki,H., Suda,T. *et al*. (2009) Craniofacial malformation in R-spondin2 knockout mice. *Biochem. Biophys. Res. Commun.*, **381**, 453–458.
36. Kim,K.A., Kakitani,M., Zhao,J., Oshima,T., Tang,T., Binnerts,M., Liu,Y., Boyle,B., Park,E., Emtage,P. *et al*. (2005) Mitogenic influence of human R-spondin1 on the intestinal epithelium. *Science*, **309**, 1256–1259.
37. Sugiura,T., Noguchi,Y., Sakurai,K. and Hattori,C. (2008) Protein phosphatase 1H, overexpressed in colon adenocarcinoma, is associated with CSE1L. *Cancer Biol. Ther.*, **7**, 285–292.
38. Tanaka,T., Ohkubo,S., Tatsuno,I. and Prives,C. (2007) hCAS/CSE1L associates with chromatin and regulates expression of select p53 target genes. *Cell*, **130**, 638–650.
39. Scherf,U., Kalab,P., Dasso,M., Pastan,I. and Brinkmann,U. (1998) The hCSE1/CAS protein is phosphorylated by HeLa extracts and MEK-1: MEK-1 phosphorylation may modulate the intracellular localization of CAS. *Biochem. Biophys. Res. Commun.*, **250**, 623–628.
40. Hu,M., Yang,J.L., Teng,H., Jia,Y.Q., Wang,R., Zhang,X.W., Wu,Y., Luo,Y., Chen,X.C., Zhang,R. *et al*. (2008) Anti-angiogenesis therapy based on the bone marrow-derived stromal cells genetically engineered to express sFlt-1 in mouse tumor model. *BMC Cancer*, **8**, 306.
41. Kumar,H., Heer,K., Greenman,J., Kerin,M.J. and Monson,J.R. (2002) Soluble FLT-1 is detectable in the sera of colorectal and breast cancer patients. *Anticancer Res.*, **22**, 1877–1880.
42. Harris,A.L., Reusch,P., Barleon,B., Hang,C., Dobbs,N. and Marme,D. (2001) Soluble Tie2 and Flt1 extracellular domains in serum of patients with renal cancer and response to antiangiogenic therapy. *Clin. Cancer Res.*, **7**, 1992–1997.
43. Moran,A.E., Hunt,D.H., Javid,S.H., Redston,M., Carothers,A.M. and Bertagnolli,M.M. (2004) Apc deficiency is associated with increased Egfr activity in the intestinal enterocytes and adenomas of C57BL/6J-Min/+ mice. *J. Biol. Chem.*, **279**, 43261–43272.
44. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al*. (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.