Statistical Analysis of Common Insertion Sites in Retroviral Insertional Mutagenesis Screens

Jeroen de Ridder[1,2], Lodewyk Wessels[1,2], Anthony Uren[2], Jaap Kool[2] & Marcel Reinders[1]

[1]Information & Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands, +31 (0)15 27 83418, J.deRidder@EWI.TUDelft.nl

[2]Department of Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

Abstract

In this paper, we propose methodologies to analyze data derived from retroviral insertional mutagenesis screens. The data has been generated by analyzing virally induced tumors in mice. The goal is to find the Common Insertion Sites (CISs), i.e. regions in the genome that have a significantly increased viral insertion rate across multiple tumors. Ideally, significance estimates of CISs should be established taking into account both the noise, arising from the random nature of the insertion process, as well as the bias, stemming from preferential insertion sites present in the genome and the data retrieval methodology.

We propose a novel method that finds CISs in a noisy and biased environment using a predefined significance level. We show that the proposed Gaussian Kernel Convolution method is flexible enough to incorporate corrections for the effects of noise and bias while maintaining the predefined significance level.

Since the locus of a CIS is frequently in the vicinity of, or within a cancer gene, detecting CISs allows for the discovery of candidate cancer genes. For this purpose a specialized MATLAB GUI visualizing the data is developed.

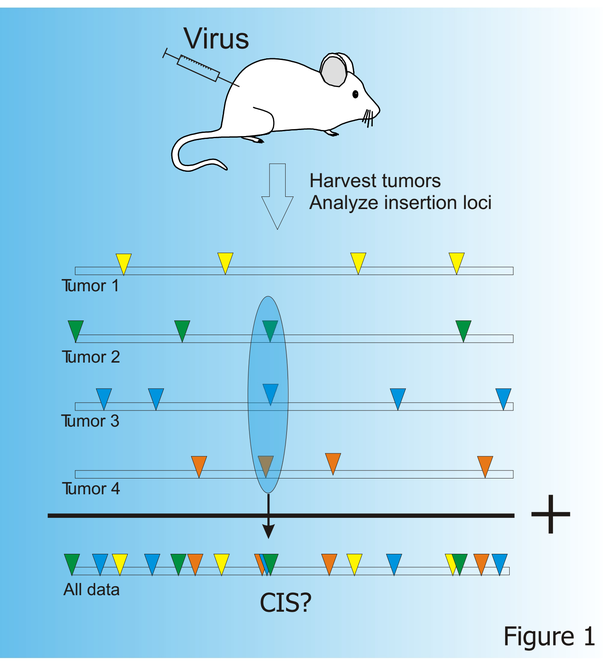

Figure 1

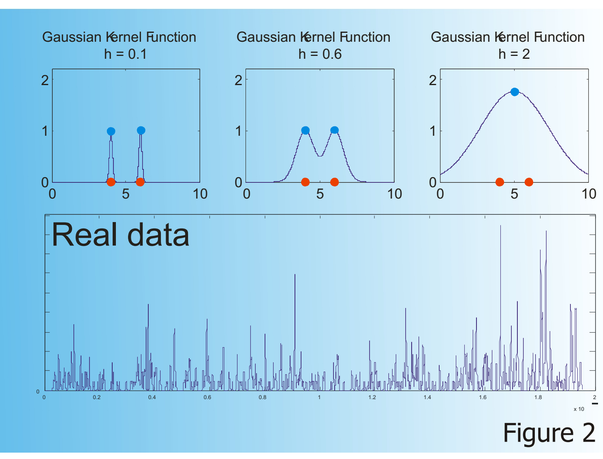

Figure 2

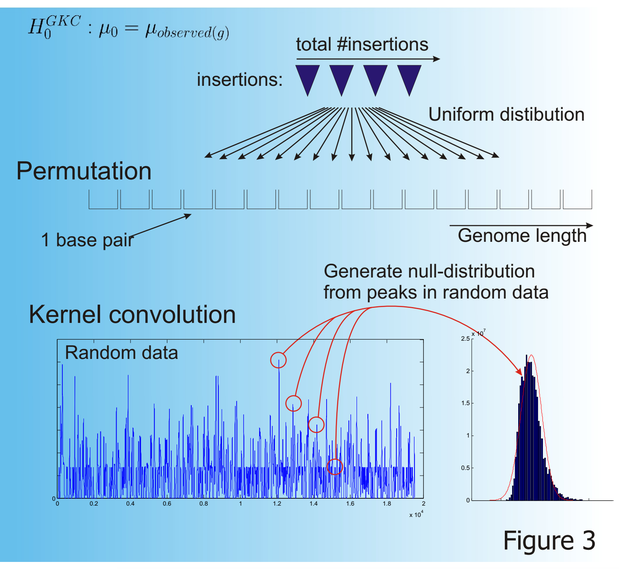

Figure 3

The data

Viral inserts may target cancer genes causing tumors to develop. In the tumor tissue one will encounter the insert that induced the proliferation far more often than inserts from non-tumor cells. Regions in the genome that are found to carry insertions in multiple independent tumors are called Common Insertion Sites (CISs). As a result, the locations of these CISs are highly correlated with the location of genes involved in tumor development.

Figure 1 shows a schematic view of the mapped data of four tumors. Note that the insertions in **different tumors** (but in the same CIS) do not need to have the exact same location since regulation of a certain cancer gene can be established over various distances. This motivates why the definition of a CIS

The problem

Find regions in the genome that have been hit by viral insertions in multiple independent tumors significantly more than expected

Methods

In this paper we introduce a Gaussian Kernel Convolution (GKC) method to estimate a density of the data.

Figure 2 shows a kernel convolution applied to two artificial insertions (red dots). When the kernel width increases (from left to right) the peaks join together. The bottom figure shows the GKC applied to real data. Finding CISs is accomplished by evaluating the significance of the heights of the peaks.

Figure 3 shows a schematic depiction of the significance analysis of the insertion data. To acquire the null-distribution the position of the N insertions are permuted. The GKC is applied to the resulting permuted insertion profile and the heights of all peaks are recorded. When this process is repeated often, a distribution of the peaks in random data results, hence the null-distribution is acquired.

Figure 4 shows the thresholded data. The threshold can be determined by considering the alpha-level in the empirical CDF of the null-distribution. This threshold can then be used to threshold the smoothed insertion estimate of the real insertion data resulting in a series of significant peaks. The Common Insertions Sites are now determined.

The definition of a CIS should be independent of its width. Therefore, it is important to consider the results of the GKC for different scale parameters (kernel widths). This gives rise to a Scale Space representation of the CISs that shows the life-span of a CIS across a range of scales. **Figure 5** gives the scale space diagram of the insertions around Myc. Vertically the scales are plotted, horizontally the position in the genome.

Software

To facilitate interactive zooming and on-the-fly adaption of the scale parameter a graphical user interface is developed. The interface displays density plots for a selected scale parameter, Common Insertion Sites and the scale space diagram. Additionally the genes are depicted, including direct links to Ensembl to provide the biologist with an easy way of browsing the insertion data and selecting candidate cancer genes. **Figure 6** shows a screen shot of the user interface.

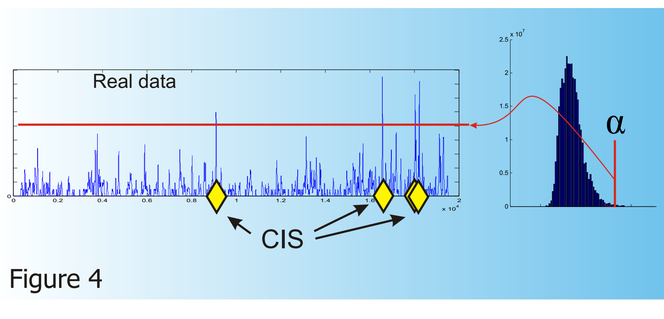

Figure 4

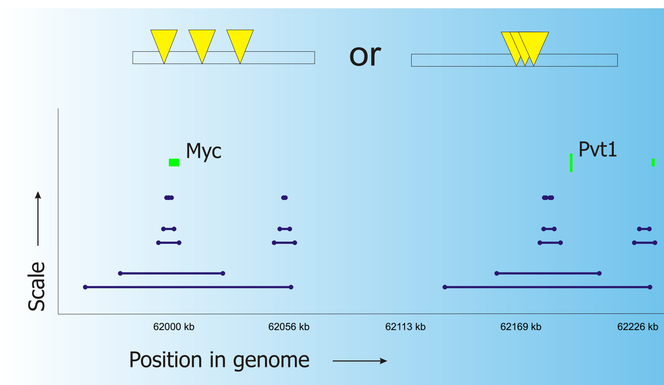

Figure 5

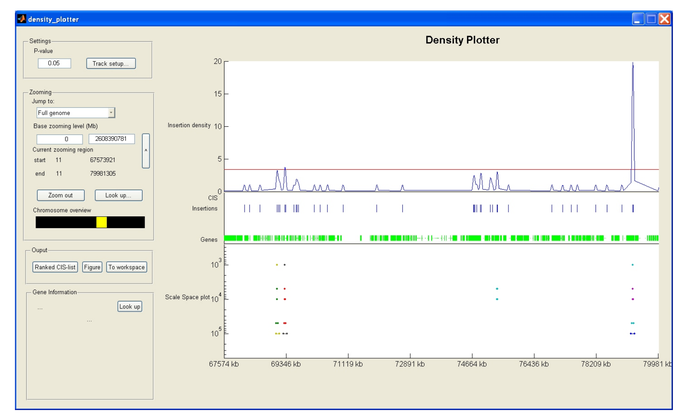

Figure 6

Conclusions

In this study we devised a method to select statistically significant Common Insertion Sites in the data stemming from Retroviral Insertional Mutagenesis Screens. The scale space approach, proposed in this study, proves to enable the detection of narrow as well as broad CISs. Together with the density plots the scale space diagrams provide a valuable visualization tool for the biologist.

References

Copeland, N. (2004). Mouse retroviral tagged cancer gene database.

Lund, A.H., Et al. (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in cdkn2a-deficient mice. Nature genetics, 32.

Mikkers, H., Et al. (2002). High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nature genetics, 32.

Mitchell, R.S., Et al. (2004). Retroviral dna integration: ASLV, HIV and MLV show distict target site preferences. PLoS, 2.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Information and Communication Theory Group

TU Delft
Delft University of Technology