Available online at www.sciencedirect.com
SCIENCE DIRECT
VIROLOGY
Virology 344 (2006) 292 – 295
www.elsevier.com/locate/yviro

Rapid Communication

# Redefining the common insertion site

Xiaolin Wu [a], Brian T. Luke [b], Shawn M. Burgess [c,*]

[a] *Laboratory of Molecular Technology, Scientific Application International Corporation-Frederick, National Cancer Institute at Frederick, NIH, Frederick, MD, USA*
[b] *Advanced Biomedical Computing Center, Scientific Application International Corporation-Frederick, National Cancer Institute at Frederick, NIH, Frederick, MD, USA*
[c] *Developmental Genomics Section, Genome Technology Branch, National Human Genome Research Institute, Bldg 50, Rm. 5537, MSC 8004, 50 South Dr., National Institutes of Health, Bethesda, MD 20892, USA*

Received 6 July 2005; returned to author for revision 10 August 2005; accepted 31 August 2005
Available online 4 November 2005

**Abstract**

Retroviral mutagenesis has been used as a powerful tool to discover genes involved in oncogenesis through a technique called Common Insertion Site (CIS) analysis where tumors are induced by proviral integrations and the genomic loci of the proviruses are identified. A fundamental assumption made in this analysis is that multiple proviral insertions in close proximity occurring more frequently than would be predicted randomly provides evidence that the genes near the integrations are involved in the formation of the tumors. We demonstrate here using data derived from MLV integrations not put under selection for tumor induction that CIS analysis as currently defined is often not a sufficient argument for a gene's significance in tumorigenesis.


*Keywords:* Insertion site; Gene; Tumorigenesis

For nearly 100 years, it has been known that viruses can be oncogenic (Ellerman and Bang, 1908; Rous, 1911). One of the best studied of such viruses is the Murine Leukemia Virus (MLV), which as its name implies, is responsible for increased numbers of cancers in the blood lineages (Friend, 1957; Gross, 1957). Recently, this property has been used in combination with high-throughput genomic approaches as a probe for genes involved in oncogenesis (Johansson et al., 2004; Kim et al., 2003; Lund et al., 2002; Mikkers et al., 2002; Shin et al., 2004; Suzuki et al., 2002). The basic approach is to infect mice with a wildtype virus that will propagate and, at a high rate, induce cancers. These tumors are isolated and the proviral integrations are mapped to genomic locations. Suzuki et al. (2002) were the first to use this approach on a large scale and they generated a model for defining what is known as a Common Insertion Site (CIS). This model compares the mapped locations of the proviruses in the isolated tumors to randomly generated integrations from 100,000 Monte Carlo trials. This allowed them to determine cutoffs for defining when two or more integrations in close proximity were significant enough to assume it didn't happen by chance (and by extension was involved in the tumorigenesis). From this analysis, they were able to develop criteria for CIS significance. Basically the cutoffs were within 30 kb for 2 integrations, 50 kb for 3 insertions or 100 kb for 4 or more integrations. Of particular note is that the criteria predicted approximately 16 false positives in the 2 insertion/30 kb cutoff (using 1200 integrations). Other criteria involved direct human interpretation of the data, but the fundamental model has been used several times since the original publication (Johansson et al., 2004; Kim et al., 2003; Lund et al., 2002; Mikkers et al., 2002; Shin et al., 2004). Given the lack of knowledge about integration sites biases for MLV, comparison to randomly generated sets of integrations was a reasonable assumption.

Recently we mapped 903 MLV integrations in HeLa cells (Wu et al., 2003). These integrations were not subject to any form of selection, so it is assumed that this data set does not have the selective bias of requiring oncogenesis used in CIS analysis. Analysis of this data set demonstrated that MLV showed a pattern of integration that was clearly not completely

* Corresponding author. Fax: +1 301 496 0474.
*E-mail address:* burgess@mail.nih.gov (S.M. Burgess).

0042-6822/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2005.08.047

Table 1
Comparison of the number of common insertion sites generated an unselected set of retroviral integrations (Unselected MLV) when compared to those generated in Suzuki et al as causative for tumorigenesis (CIS) and their random integration model (Random)

| | Random[a] | Unselected MLV[b] | CIS[c] |
|---|---|---|---|
| 2 integrations | 16.3 | 59.8 | 95 |
| 3 integrations | 0.3 | 4.0 | 21 |
| 4+ integrations | 0.4 | 4.0 (0) | 28 (17) |

Particularly in the 2 integration category, there is an enrichment for CIS occurrences in the unselected integrations when compared to random.

[a] Number taken from Suzuki et al. (web supplement) based on 100,000 Monte Carlo simulations.

[b] Based on data from Wu et al. Number is normalized to the Suzuki et al. data by the formula: ($n$/903)1200, where $n$ = the number of CIS based on criteria from Suzuki et al. Number in parentheses is the number of CIS with greater than 4 integrations.

[c] From Suzuki et al. Based on 1200 integrations. Number in parentheses is the number of CIS with greater than 4 integrations.

random. Approximately 20% of the integrations occurred within ±5 kb of the transcriptional start site (Refseq genes, Nov. 2002, UCSC human genome Hg13), where in a randomly generated set that frequency was approximately 4% ($P < 0.0001$). In terms of the null hypothesis for CIS analysis this is problematic. The criteria CIS's are based on a comparison to a random set, but there is a clear bias in integration sites even without selection, therefore the null hypothesis for CIS analysis must now take this non-random behavior into account.

To establish how much of a problem the intrinsic biases of MLV site selection to CIS analysis are, we remapped the MLV integrations to the most current build of the human genome (May 2004, UCSC human genome Hg17). We then applied the original criteria for CIS analysis, ignoring any CIS determined by human interpretation (i.e. integrations near a known oncogene) and then normalized the data to the number of integrations mapped by Suzuki et al. The data are summarized in Table 1 (see supplemental data for precise coordinates for all the CIS). The major feature of this analysis is the relatively high incidence of CIS sites in the unselected MLV data, particularly when the criterion is two integrations within 30 kb. Using this definition, nearly two thirds of the 2-integration CIS's (59.8 of 95) in Suzuki et al. can be explained by the null hypothesis of natural retroviral site biases. When looking at 3 integration CIS's nearly 20% (4 of 21) integrations can again be explained by the null hypothesis. Once the integration number in the CIS is over 4, an interesting division occurs. After normalization in the unselected set of MLV integrations, there were 4 CIS with 4 integrations and 0 CIS with more than 4 integrations. In Suzuki et al., there were 11 CIS with 4 integrations, 17 CIS with more than 4 integrations, and one CIS having 55 integrations (Sox4). This would suggest that CIS with 3 or 4 integrations would still need an additional level of proof with more than a 20% chance of a false positive, while integration frequencies above 4 would be extremely rare in the null hypothesis of no selection.

One argument that has been used to demonstrate the efficacy of CIS analysis is that many genes identified by the analysis had already been demonstrated to have a role in oncogenesis using other methods. We determined the closest gene for each CIS in the unselected MLV data. We selected only the genes that have at least one publication associated with them and searched PubMed for the name of the gene in combination with the word cancer. Of the 38 genes tested in this fashion, 14 (37%) of them could be demonstrated to be positively associated with cancer using only this simple and far from comprehensive criterion (see supplemental data). This small exercise demonstrates that a pre-established role in cancer is not sufficient support for the efficacy of the CIS technique. We previously demonstrated that MLV integrations are biased towards genes with higher expression levels (Wu et al., 2003) and the integration mapping was done in the tumor derived HeLa cells. Thus, it can be argued that MLV is biasing integrations into locations that are transcriptionally more active because the genes are related to cancer formation. Therefore, it may be unclear if the integrations are causative for the tumor or are targeted to that location because it is a tumor.

We developed a simple model to simulate the integration biases we had established in our original paper. Instead of simulating completely random integration, we limited 25%

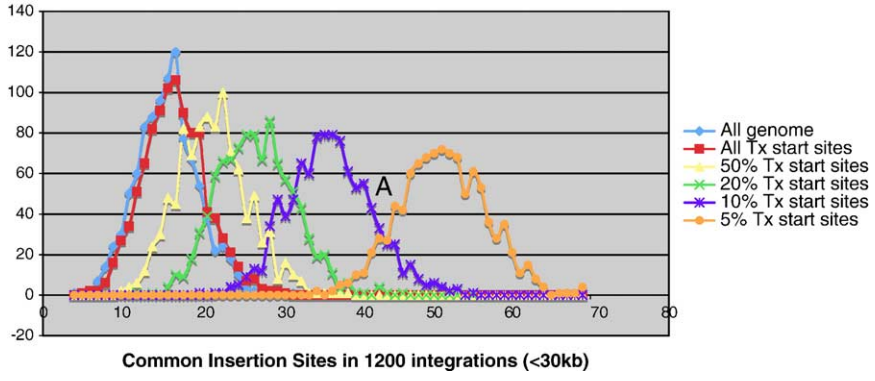


Fig. 1. The frequency of retroviral integrations creating common insertion sites (CIS) using various models of integration and repeated 1000 times for each model. In Wu et al., 25% of the integrations were shown to land ±5 kb of the transcriptional start site, and 75% were considered equivalent to random based on various genomic features. The model allows 75% of the integrations to occur randomly and 25% to integrate in a Poisson distribution ±5 kb around the transcriptional start site. The simulation was then tested with all genes as targets, or increasingly smaller numbers of genes available for integration (to imitate biases for level of gene expression). Not until the genes are limited to 5% of all genes are CIS generated at a rate near that of our experimental data.

(percentage found in our calculations using the most recent human genome freeze) of the integrations to a Gaussian distribution ±5 kb around the transcriptional start site of RefSeq genes. The remaining 75% of the integrations were allowed to integrate randomly (Fig. 1). We then ran the simulation 1000 times on the mouse genome (MM5) with 1200 integrations in each simulation to match the number of integrations from Suzuki et al. This model was indistinguishable from random integration in terms of CIS frequency (Fig. 1 red and blue lines). MLV also has a bias towards more highly expressed genes. As a way to simulate restricted integration based on the level of gene expression in addition to the preference for the 5′ end of genes, we used the simple model of considering genes to be either "on" or "off" and only allowed the 25% category of integrations in genes to ones that were "on". We tested models that used 1/2, 1/5, 1/10, and 1/20 of the total 18,366 RefSeq genes in the current mouse genome build MM5 (Fig. 1). It was not until we limited the model to 1/20 of the RefSeq genes that we were able to demonstrate CIS frequencies similar to those seen in our data set. The average number of CIS was approximately 55, with the highest seen being 69 CIS. We examined our original integration data to determine how many of the CIS in our unselected viral integrations had at least one integration within ±5 kb of the transcription start site. Only 15 of the 68 CIS satisfied this criterion. This was a strong indication that the 5′ bias and gene expression were not the only factors influencing integration position.

There are two important ramifications that come from our analyses. The first is that there appears to be additional factors influencing MLV site selection beyond the previously described affinity for the 5′ ends of genes. Our modeling demonstrates that such a bias is insufficient to explain the frequency of CIS in our unselected integrations without an extreme limitation in the number of genes available for retroviral integration (1/20 of all genes). As only 15 of the 68 CIS in our data have even one integration near the 5′ end of a gene, there must be other aspects of the chromatin influencing proviral integration. We do not currently know what features of chromatin are influencing these chromosomal integration hotspots. One possible way of thinking about the observed bias towards the 5′ ends of genes is that the transcriptional start regions of genes have structural properties (e.g., open chromatin) that are similar to other regions of the chromatin that also have a propensity for retroviral integration. Thus, the 25% of integrations that appear to be biasing towards the 5′ ends of genes are merely a subset of the total global chromatin features that are considered preferential for MLV integration. Second, the traditional definition of a common insertion site being statistically significant for tumorigenesis can no longer be used and a new definition must take into account the natural biases of whatever virus is being used. The data from our unselected integrations generates CIS at a frequency high enough to create problems with interpretation based on the traditional definition. Nearly 2/3 of the CIS that consist of 2 integrations within 30 kb could be accounted for by the null hypothesis of no selection and 20% of the 3 or 4 integration CIS's. Only when integration frequencies get above 4 integrations in a CIS (from a data set of around 1200 integrations) can you definitively say that the CIS is highly unlikely to occur by chance. It is quite likely, even probable that the CIS of 4 and below are significant in tumorigenesis, but they would require a much higher level of proof to their significance than what is supplied by CIS analysis, as they could realistically have occurred by chance.

In conclusion, CIS analysis is a very powerful tool for identifying candidate genes involved in oncogenesis as well as normal development, but the statistical analysis as it is currently done does not take into account the natural integration preferences of the retroviruses being used. These biases have both the natural tendency to generate CIS's as well as to integrate into genes that are potentially relevant to cancer. We propose a much more stringent criteria based on modeling the unselected integration patterns of whatever retrovirus is being used, and comparing that unselected model to the observed integration frequencies in the tumors. For MLV integrations, CIS of 5 or more integrations carry statistical significance in a set of 1200, while below 5 integrations have an increased possibility of occurring by chance. It is not clear what the implications of this finding are for the potential increased risk MLV based vectors have when used for gene therapy. The indications are clear that there are genomic regions that are preferentially targeted, but why or how those regions are targeted is still unknown.

## Acknowledgments

We would like to thank B. Westermark for asking the initial question that prompted this research and N. Copeland for critical reading of the manuscript. This project has been funded in whole or in part with Federal Funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400 (Article H.36 of the Prime Contract) and the intramural program of the National Human Genome Research Institute. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.virol.2005.08.047.

## References

Ellerman, V., Bang, O., 1908. Experimentelle Leukämie bei Hühnern. Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg. Abt. Orig. 46, 595–609.

Friend, C., 1957. Cell-free transmission in adult Swiss mice of a disease having the character of a leukemia. J. Exp. Med. 105, 307–319.

Gross, L., 1957. Development and serial cell-free passage of a highly potent strain of mouse leukemia virus. Proc. Soc. Exp. Biol. Med. 94, 767–771.

Johansson, F.K., Brodd, J., Eklof, C., Ferletta, M., Hesselager, G., Tiger, C.F., Uhrbom, L., Westermark, B., 2004. Identification of candidate cancer-

causing genes in mouse brain tumors by retroviral tagging. Proc. Natl. Acad. Sci. U.S.A. 101 (31), 11334–11337.

Kim, R., Trubetskoy, A., Suzuki, T., Jenkins, N.A., Copeland, N.G., Lenz, J., 2003. Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. J. Virol. 77 (3), 2056–2062.

Lund, A.H., Turner, G., Trubetskoy, A., Verhoeven, E., Wientjens, E., Hulsman, D., Russell, R., DePinho, R.A., Lenz, J., van Lohuizen, M., 2002. Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat. Genet. 32 (1), 160–165.

Mikkers, H., Allen, J., Knipscheer, P., Romeijn, L., Hart, A., Vink, E., Berns, A., 2002. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat. Genet. 32 (1), 153–159.

Rous, P., 1911. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. J. Exp. Med. 13, 397–411.

Shin, M.S., Fredrickson, T.N., Hartley, J.W., Suzuki, T., Agaki, K., Morse III, H.C., 2004. High-throughput retroviral tagging for identification of genes involved in initiation and progression of mouse splenic marginal zone lymphomas. Cancer Res. 64 (13), 4419–4427.

Suzuki, T., Shen, H., Akagi, K., Morse, H.C., Malley, J.D., Naiman, D.Q., Jenkins, N.A., Copeland, N.G., 2002. New genes involved in cancer identified by retroviral tagging. Nat. Genet. 32 (1), 166–174.

Wu, X., Li, Y., Crise, B., Burgess, S.M., 2003. Transcription start regions in the human genome are favored targets for MLV integration. Science 300 (5626), 1749–1751.