OPEN ACCESS Freely available online

PLoS one

Real-Time Definition of Non-Randomness in the Distribution of Genomic Events

Ulrich Abel[1,2], Annette Deichmann[1], Cynthia Bartholomae[1], Kerstin Schwarzwaelder[1], Hanno Glimm[1], Steven Howe[2], Adrian Thrasher[3,4], Alexandrine Garrigue[5], Salima Hacein-Bey-Abina[5,6], Marina Cavazzana-Calvo[5,6], Alain Fischer[5,7], Dirk Jaeger[1], Christof von Kalle[1,8]*, Manfred Schmidt[1]*

1 Department of Translational Oncology, National Center for Tumor Diseases, Heidelberg, Germany, **2** Department of Medical Biostatistics, Tumor Center Heidelberg-Mannheim, Heidelberg, Germany, **3** Molecular Immunology Unit, Institute of Child Health, University College London, London, United Kingdom, **4** Department of Clinical Immunology, Great Ormond Street Hospital NHS Trust, London, United Kingdom, **5** INSERM Unit 768, Hôpital Necker and Faculté de Médecine Université René Descartes Paris V., Paris, France, **6** Département de Biothérapies, Hôpital Necker, Paris, France, **7** Unité d'Immunologie et d'Hématologie Pédiatriques, Hôpital Necker, Paris, France, **8** Division of Experimental Hematology, Cincinnati Childrens Research Foundation, Cincinnati, Ohio, United States of America

Features such as mutations or structural characteristics can be non-randomly or non-uniformly distributed within a genome. So far, computer simulations were required for statistical inferences on the distribution of sequence motifs. Here, we show that these analyses are possible using an analytical, mathematical approach. For the assessment of non-randomness, our calculations only require information including genome size, number of (sampled) sequence motifs and distance parameters. We have developed computer programs evaluating our analytical formulas for the real-time determination of expected values and *p*-values. This approach permits a flexible cluster definition that can be applied to most effectively identify non-random or non-uniform sequence motif distribution. As an example, we show the effectivity and reliability of our mathematical approach in clinical retroviral vector integration site distribution.

Citation: Abel U, Deichmann A, Bartholomae C, Schwarzwaelder K, Glimm H, et al (2007) Real-Time Definition of Non-Randomness in the Distribution of Genomic Events. PLoS ONE 2(6): e570. doi:10.1371/journal.pone.0000570

INTRODUCTION

With the sequences of complete genomes available [1–4], and accelerating technologies for high-throughput sequencing [5] genome wide sequence analyses of individual samples will soon become reality. Comparative analyses of sequence composition and sequence motif distribution have become central parts of genome and transcriptome research, providing new insights on evolution, physiology and medical diagnosis [6–15]. Our understanding of integrating viruses and related vectors in gene therapy trials is an interesting example of such approaches. Since the completion of the human and murine genome sequencing projects the location of the vector in the cellular genome can be defined precisely, allowing the determination of possible vector integration induced effects on the surrounding genomic DNA regions at the molecular level. Integration site analyses have gained increasing interest with the dramatic development of a retroviral vector-induced lymphoproliferative disease in 3 patients cured of X-linked severe combined immunodeficiency (X-SCID) that was triggered by insertional activation of the proto-oncogene *LMO2* [16,17]. Meanwhile, insertion induced side effects have been identified ranging from immortalization [18] to clonal dominance [19–22] and even oncogenesis [23–25] in a variety of gene therapy studies. These studies have in common that a clustering of integration sites (IS) in certain genomic loci was detectable, and likely provided a selective advantage for the affected cell clone.

The clustering of integrations, termed common integration sites (CIS), as an indicator for clone selection has already been used in concerted retrovirus insertional mutagenesis studies that aimed to identify new cancer genes by determining the gene configuration near frequently affected integration site loci [26–28]. For CIS determination, computer simulations were performed to assess non-randomness of IS distribution in tumors [28]. To validate the correctness of our mathematical approach defining non-randomness and non-uniform sequence motif distribution, we analyzed the IS distribution and presence of CIS in 2 successful clinical SCID-X1 studies [29,30, unpublished data]. We considered 2, 3 or 4 insertions as CIS of 2^{nd}, of 3^{rd} or 4^{th} order if they fell within a 30 kb, 50 kb or 100 kb window of genomic sequence from each other, respectively. Simultaneously, we performed computer simulations written in open source 'R'-language (http://cran.r-project.org) for which a window of size d_n (d_n = the maximum distance defining a CIS of order n) was shifted through the ordered sequence of the IS. For each window $W(j) = [IS(j), IS(j)+d_n]$ it was then counted how many CIS of order n including $IS(j)$ as first element were contained in $W(j)$. We show that our mathematical approach for defining biased IS distribution is comparable to the output of computational simulations. It may have advantages in performance of large quantities of individual analyses. Even if the null hypothesis of random uniform allocation is not adequate, as it is known from retroviral vector integration [31], our calculations can address segments of the genome located between sites of predilection for virus integration and can be extended to address non-uniform sequence motif distributions.

Academic Editor: Xiaolin Wu, National Cancer Institute at Frederick, United States of America

Received February 23, 2007; **Accepted** June 6, 2007; **Published** June 27, 2007

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (Grant SPP1230), the German Ministry of Education and Research (Grant TREATID) and the EU (VIth Framework Program, CONSERT and CLINIGENE).

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: christof.kalle@nct-heidelberg.de (CK); manfred.schmidt@nct-heidelberg.de (MS)

RESULTS AND DISCUSSION

Part 1: Random uniform allocation of IS

For the purpose of this discussion, the unit of observation (location and distance) is kilobasepair (kb). We assume that a number n_{is} of IS is randomly allocated (with a uniform distribution) to the locations of a genome consisting of g kb. A CIS of order n is an n-tuple of IS such that the maximum distance between the lowest and highest position is no greater than a fixed bound.

Further terminology

- d_n, defining "size" or distance of a CIS of order n, i.e. maximum permissible distance between any two members of a CIS of order n.;
- P_n, probability that a given (sub)set of n IS that are randomly allocated form a CIS of order n
- $P(m,d)$, probability that a given subset of m randomly allocated IS has a span (= maximum distance between any two elements) of exactly d.
- E_n, expected value of the number of CIS of order n

We start with the elementary observation that E_n equals P_n times the number of subsets of IS consisting of n elements:

$$E_n = \binom{n_{is}}{n} \cdot P_n \tag{1}$$

Clearly,

$$P_n = \sum_{d=0}^{d_n} P(n,d) \tag{2}$$

It remains to determine $P(n,d)$. First note that $P(1,d) = 0$ for $d>0$. Furthermore, for all $m \geq 1$:

$$P(m,0) = \frac{1}{g^{m-1}} \tag{3}$$

A recursive formula for $P(m,d)$, $d>0$, can be derived by breaking down the potential CIS of order m into subsets of m–1 elements having a span of $d' \leq d$, to which an m-th IS is added such that the maximum span is exactly d:

$$P(m,d) = \frac{1}{g}\{\sum_{d'=0}^{d-1} [2 \cdot P(m-1,d')] + (d+1)P(m-1,d)\} + r \tag{4}$$

where r is a negligible correction term that arises because the uncorrected recursion formula is strictly valid only for subsets of IS that have a distance $\geq d$ from the telomeres.

By mounting the recursive ladder ($m = 1,\ldots,n$), these formulas successively yield $P(n,d)$, P_n, and E_n. In particular, one easily obtains ($d>0$):

$$P(2,d) \approx \frac{2}{g}$$

$$P(3,d) \approx \frac{6d}{g^2}$$

$$P(4,d) \approx \frac{12d^2+2}{g^3}$$

Plugging this into equations (2) and (1) yields for the expected value E_n:

$$E_2 \approx \binom{n_{is}}{2} \frac{(2d_2+1)}{g}$$

$$E_3 \approx \binom{n_{is}}{3} \frac{\{3d_3(d_3+1)+1\}}{g^2}$$

$$E_4 \approx \binom{n_{is}}{4} \frac{1+2d_4\{1+(d_4+1)(2d_4+1)\}}{g^3}$$

As shown in **Table 1**, our mathematical approximation corresponds extremely well to the mean values found in 50000 simulation runs.

Statistical inferences, such as the calculation of p-values, can be based on the observation that, under the null hypothesis (H_0) of random uniform allocation of the IS, the number of CIS of order n is (approximately) Poisson distributed with parameter $\lambda = E_n$. Thus, if the random variable X denotes the number of CIS of order n, and $X = k$ is observed in a trial, then the p-value $P(X \geq k)$ of this observation calculated under H_0, i.e. from the Poisson distribution $P_o(E_n)$, is given by

$$P(X \geq k | H_o) = 1 - \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} e^{-\lambda} = P(\chi^2 \leq 2E_n),$$

where the random variable χ^2 has a chi-square distribution with 2 k degrees of freedom [32,33].

The Poisson approximation to the true random distribution of CIS is exceedingly close. In fact, if the number of simulation runs is sufficiently high, the simulated distribution is virtually undistinguishable from $P_o(E_n)$. In particular, both the expected values and the p-values derived from $P_o(E_n)$ are nearly identical to those obtained in computer simulations. The latter point is apparent from **Table 2**, where for a final proof of principle of our mathematical calculations, results of the analysis of our integration data set retrieved from two clinical SCID-X1 therapy trials [unpublished data] are given.

The p-value can be calculated by means of either of the following commands ('R' code): 1–ppois(lambda = E_n, $q = k$–1) or pchisq(df = $2k$, $q = 2E_n$). Using the data of **Table 2** (first line) 1–ppois(lambda = 0.19, $q = 2$) or pchisq(df = 6, $q = 0.38$). In both

Table 1 Mean values for random CIS formation (1000 IS) determined either with computer simulations or mathematically.

Order of CIS	Mean Value Mathematical Formula	Mean Value Computer Simulations
2nd	$E_2 \approx \binom{1000}{2} \frac{61}{3.12 \cdot 10^6} = 9.77$	9.75
3rd	$E_3 \approx \binom{1000}{3} \frac{7651}{(3.12 \cdot 10^6)^2} = 0.13$	0.13
4th	$E_4 \approx \binom{1000}{4} \frac{4.06 \cdot 10^6}{(3.12 \cdot 10^6)^3} = 0.01$	0.01

Simulations were performed with 50000 runs each. g, haploid size of the human genome: 3.12 x 10^6 kb; d_n, genomic window size [kb] for CIS of n^{th} order: $d_2 = 30$, $d_3 = 50$, and $d_4 = 100$; n_{is}, number of (assumed) sampled integration sites: 1000.

doi:10.1371/journal.pone.0000570.t001

Table 2 Comparative analysis of mean values and *p*-values obtained computationally ('Simulation') or mathematically ('Formula').

CIS	IS	MV Simulation	MV Formula	*p*-Value Simulation	*p*-Value Formula
3	140	0.188	0.190	0.0009	0.001
1	134	0.175	0.174	0.16	0.16
4	102	0.100	0.101	0	3.9×10^{-6}
15	304	0.899	0.900	0	6.8×10^{-14}
102	572	3.200	3.193	0	$<10^{-16}$

The results refer to the presence of CIS detected in 2 clinical X-SCID gene therapy studies [unpublished data]. Simulations were performed with 50000 runs on the haploid size of the human genome (3.12×10^6 kb). *P*-values estimated from simulations equal the proportion per 50000 runs in which the number of CIS was at least as high as the number observed in the trials. The genomic window size chosen for CIS of 2nd order was 30kb. CIS, number of identified CIS of 2nd order in patient and control samples pre- and post-transplant; IS, number of all unique identified integration sites in patient and control samples pre- and post-transplant; MV, mean value.

doi:10.1371/journal.pone.0000570.t002

instances, the result is 0.00099. Alternatively, the table of the chisquare distribution with 6 degrees of freedom can be used to look up the probability $P(X\leq0.38)$. One should note that, for low E_n, the *p*-value of a single observed CIS is virtually identical to E_n. This implies that, for $n>5$, no *p*-values need to be calculated (and hence no formulas are required for E_n, $n>5$), because even with an extremely liberal definition of the CIS ($d_5 = 500$) and a fairly high number of IS ($n_{is} = 1000$) a single CIS of order 5 will be statistically significant ($p = 0.027$).

Part 2: Non-uniform allocation of IS

Defining non-randomness in the clustering of genomic events often requires additional precautions as sequence structures of interest may already have known specific distribution biases. In the case of our clinical example (unpublished data), it is known that retroviral vectors based on the murine leukaemia virus (MLV) tend to integrate into gene coding regions preferentially near the transcriptional start site (TSS) [34-36]. It is also proposed that additional factors, indeed mostly unknown, may influence the accessibility of vectors to certain genomic DNA regions [37]. Thus, the null hypothesis of random uniform allocation of MLV IS distribution may not be adequate according to the current 'state of the art', as has recently been argued [31]. In line with this study, we portioned the genome into 2 adequate areas that differ in the likelihood of getting targeted by vectors.

Further terminology

n_{TSS},	number of TSS;
T5,	an interval of +/-5kb around a TSS;
GT5,	union of all T5
$n_{is,Mix}$, $n_{is,Comp}$,	number of IS occuring in GT5 and in the complement of GT5, respectively
$n_{cis,GT5}$, $n_{cis,Mix}$, $n_{cis,Comp}$,	number of CIS occurring in GT5, both in GT5 and in the complement of GT5 and in the complement of GT5 only, respectively.

Clearly, the expected value E_n of the number CIS of order n is given by the following sum:

$$E_n = E(n_{cis,GT5}) + E(n_{cis,Mix}) + E(n_{cis,Comp}) \quad (5)$$

In the following it will be shown how to calculate the terms on the right side of (5). We start with the expected value of $n_{cis,GT5}$ fore what we assume that vector integration into any T5 occurs with the same probability. Then

$$E(n_{cis,GT5}) = n_{TSS}\cdot E(X), \quad (6)$$

where X is the number of CIS (among those occurring in GT5) that occur in a fixed T5. Observing that i IS in a fixed T5 yield $\binom{i}{n}$ CIS of order n in this T5 one easily obtains the expected value of X

$$E(X) = P(X=n)\cdot1 + P(X=n+1)\binom{n+1}{n} + P(X=n+2)\binom{n+2}{n} + \ldots \quad (7)$$

Since X is binomially distributed as $\sim B(n_{is},GT5,1/n_{TSS})$,

$$P(X=i) = \binom{n_{is,GU5}}{i}\left(\frac{1}{n_{TSS}}\right)^i\left(\frac{n_{TSS}-1}{n_{TSS}}\right)^{n_{is,GU5}} \quad (8)$$

Merging equations (6)–(8) yields the desired formula for $E(n_{cis,GT5})$:

$$E(n_{cis,GU5}) = n_{TSS}\sum_{i=n}^{n_{is,GU5}}\binom{n_{is,GU5}}{i}\binom{i}{n}\left(\frac{1}{n_{TSS}}\right)^i\left(\frac{n_{TSS}-1}{n_{TSS}}\right)^{n_{is,GU5}} \quad (9)$$

If $n_{is,GT5}$ is small compared to n_{TSS} (undoubtedly, this is mostly the case), terms of higher order can be neglected so that, because $(n_{TSS}-1)/n_{TSS}\approx1$, formula (9) simplifies to

$$E(n_{cis,GU5}) \approx n_{TSS}\binom{n_{is,GU5}}{n}\left(\frac{1}{n_{TSS}}\right)^n = \binom{n_{is,GU5}}{n}\left(\frac{1}{n_{TSS}}\right)^{n-1} \quad (10)$$

Notice that formulas (6)–(10) do not depend on the spatial distribution of the IS within the T5. (It is unnecessary to account for the closeness of IS within T5 because *any* pair – or triple, quadruple etc., for that matter – of IS within a T5 yields a CIS.)

Clearly, the expected value of $n_{cis,Mix}$ $E(n_{cis,Mix})$ is not independent of the distance between the IS and the TSS. Thus, inevitably, assumptions regarding the spatial distribution for the IS will influence its value. In the sequel, a formula for $E(n_{cis,Mix})$ shall be derived for the case $n = 2$. As before, CIS of order 2 are defined by a maximum distance d_2 of 30kb between the IS.

If the TSS are indistiguishable with respect to the probability distribution of the integrations, then

$$E(n_{cis,Mix}) = n_{is,GT5}\cdot n_{is,Comp}\cdot n_{TSS}\cdot p_{Mix}, \quad (11)$$

where p_{Mix} denotes the probability that an arbitrary pair of IS (with one element in GT5 and one element in the complement of GT5) forms a CIS of order 2 around a fixed TSS.

We will assume that the distributions of IS within a T5 and within +/-35 kb around a TSS are symmetric. Then, again using kb as unit of distance,

$$p_{Mix} = 2\int_{x=0}^{10} f(x)\int_{y=x-30}^{0} g(y)\,dy\,dx. \quad (12)$$

In formula (12) the points $x = 0$ and $y = 0$ correspond to the TSS-5; $f(x)$ designates the probability density function of vector

integrations in T5; and $g(y)$ designates the corresponding density function in [TSS-35, TSS-5].

Formula (12) shall be evaluated for two special cases:

Case 1: Vector integrations are uniformly distributed in GT5 and in the complement of GT5, respectively. I.e.,

$$f(x) \equiv 1/(n_{\mathrm{TSS}} \cdot 10)$$

$$g(x) \equiv 1/(g - n_{\mathrm{TSS}} \cdot 10).$$

Solving the integrals in formula (12) we have

$$p_{Mix} = \frac{400}{10 n_{TSS}(g - 10 n_{TSS})} \tag{13}$$

Case 2: As above, vector integrations in the complement of GT5 are assumed to be uniformly distributed. However, a triangular distribution is assumed for $f(x)$. The corresponding formula is easily calculated:

$$f(x) = \begin{cases} x/(25 n_{TSS}) & if \quad x \le 5 \\ (2/5 - x/25)/n_{TSS} & if \quad x \ge 5 \end{cases}$$

By plugging this into (12) we get

$$p_{Mix} = \frac{170}{3 n_{TSS}(g - 10 n_{TSS})} \tag{14}$$

It may be surprising that a triangular distribution in T5 results in a higher expected value for $n_{\mathrm{cis,Mix}}$ than a uniform distribution. However, this becomes more plausible if one notes that a higher value is also obtained if the IS are concentrated in an extreme manner within the T5, viz. in a one-point distribution with total mass in the TSS. In this special case (which is particularly easy to evaluate), $p_{\mathrm{Mix}} = 50/(n_{\mathrm{TSS}}(g - n_{\mathrm{TSS}}))$.

If, with respect to the formation of CIS, the complement of GT5 could be regarded as a continuum, the expected value of $n_{\mathrm{cis,Comp}}$ would be given by the formulas developed in **Part 1** of this contribution. In the case of retroviral (MLV) vectors, however, the complement of GT5 has rather to be viewed as a partitioned set consisting of approximately TSS disjoint intervals. It follows that that the residual term on the right-hand side of equation (4) (**Part 1**) may no longer be negligible. Note however, the assumption of a continuum clearly tends to lead to an overestimation of the number of CIS, because the boundaries of the components reduce the number of CIS occurring in their neighborhood. It follows that the formulas derived in **Part 1** form an upper bound for $E(n_{\mathrm{cis,Comp}})$. In particular, the true p-values are less or equal to the values calculated by means of the formulas derived in **Part 1**.

Table 3 Formulas based statistical analysis of the results on CIS formation in clinical samples derived from 2 clinical X-SCID gene therapy studies [unpublished data].

CIS	IS	MV Uniform*	MV Triangular§	*p*-Value Uniform*	*p*-Value Triangular§
3	140	0.191	0.212	0.001	0.0014
1	134	0.175	0.195	0.161	0.177
4	102	0.101	0.124	4.0×10^{-6}	6.1×10^{-6}
15	304	0.905	1.006	7.4×10^{-14}	3.3×10^{-13}
102	572	3.212	3.568	$<10^{-16}$	$<10^{-16}$

Calculations were performed on the haploid size of the human genome (3.12×10^6 kb) and on the basis of an IS skewing (25% of all IS) to the +/− 5 kb TSS region, for which an (*) uniform or a (§) triangular IS distribution, respectively, was assumed. 75% of IS were assumed to be uniformly distributed over the remaining human genome. The genomic window size chosen for CIS of 2nd order was 30 kb. CIS, number of identified CIS of 2nd order in patient and control samples pre- and post-transplant; IS, number of all unique identified integration sites in patient and control samples pre- and post-transplant; MV, mean value.

doi:10.1371/journal.pone.0000570.t003

Therefore, any positive statements regarding statistical significance remain valid. Moreover, the overestimation is probably fairly small given that the sections of GT5 located between the TSS are mostly rather wide compared to the length defining a CIS.

Indeed, the null hypothesis of non-uniform allocation for IS distribution does not substantially change the results we have obtained based on the hypothesis of a random uniform allocation for CIS formation in our clinical samples (**Table 2**), as is shown in **Table 3**.

Our mathematical formulas allow a reliable, straightforward calculation of non-randomness in CIS and other genomic event distributions under the null hypothesis of uniform and non-uniform allocation. Using formula based workspaces (available on request), expected values and p-values can be calculated with ease in real-time. They may be preferable to computer simulations when (routine) high-speed processing of large quantities of analyses is needed. Our approach enables a closely problem-oriented, highly exact evaluation of non-randomness that is useful for assessing IS distribution in clinical trials and for assessing the distribution of any sequence motif of interest in a natural or artificial genome.

ACKNOWLEDGMENTS

Author Contributions

Conceived and designed the experiments: MS Cv UA. Performed the experiments: AD Cb KS UA. Analyzed the data: MS AD Cb KS HG UA. Contributed reagents/materials/analysis tools: AF AT DJ SH AG SH MC. Wrote the paper: MS Cv UA. Other: Senior authors: Cv MS.

REFERENCES

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287: 2185–2195.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
3. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129–149.
4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
6. Camargo AA, Samara HP, Dias-Neto E, Simao DF, Bigotto IA (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 98: 12103–12108.
7. Riva A, Delorme M-O, Chevalier T, Guilhot N, Henaut C, et al. (2004) The difficult interpretation of transcriptome data: the case of the GATC regulatory network. Computational Biology and Chemistry 28: 109–118.
8. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, et al. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14: 2121–2127.

9. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40–45.
10. Wang J, Song L, Gonder MK, Azrak S, Ray DA, et al. (2006) Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. Gene 365: 11–20.
11. Garrigan D, Hammer MF (2006) Reconstructing human origins in the genomic era. Nat Rev Genet 7: 669–680.
12. Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, et al. (2003) Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics 19: 549–552.
13. Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromosomes and potential role in higher order chromatin organization and function. Bioinformatics 19: 681–685.
14. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, et al. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 126: 1203–1217.
15. Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 18: 1803–1818.
16. Hacein-Bey-Abina S, von Kalle C, Schmidt M, McCormack MP, Wulffraat N, et al. (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science 302: 415–419.
17. Hacein-Bey-Abina S, von Kalle C, Schmidt M, Le Deist F, Wulffraat N, et al. (2003) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. N Engl J Med 348: 255–256.
18. Du Y, Jenkins NA, Copeland NG (2005) Insertional mutagenesis identifies genes that promote the immortalization of primary bone marrow progenitor cells. Blood 106: 3932–3839.
19. Hematti P, Hong BK, Ferguson C, Adler R, Hanawa H, et al. (2004) Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. PLoS Biology 2: e423.
20. Calmels B, Ferguson C, Laukkanen MO, Adler R, Faulhaber M, et al. (2005) Recurrent retroviral vector integration at the MDS1-EVI1 locus in non-human primate hematopoietic cells. Blood 106: 2530–2533.
21. Kustikova O, Fehse B, Modlich U, Yang M, Dullmann J, et al. (2005) Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science 308: 1171–1174.
22. Ott MG, Schmidt M, Schwarzwaelder K, Stein S, Siler U, et al. (2006) Correction of X-linked chronic granulomatous disease by gene therapy is augmented by insertional activation of MDS/EVI1, PRDM16 or SETBP1. Nat Med 12: 401–409.
23. Li X, Düllmann J, Schiedlmeier B, Schmidt M, von Kalle C, et al. (2002) Murine leukemia induced by retroviral gene marking. Science 296: 497.
24. Modlich U, Kustikova OS, Schmidt M, Rudolph C, Meyer J, et al. (2005) Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood 105: 4235–4246.
25. Montini E, Cesana D, Schmidt M, Sancito F, Ponzoni M, et al. (2006) Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat Biotechnol 24: 687–696.
26. Mikkers H, Allen J, Knipscheer P, Romeijn L, Hart A, et al. (2002) High-throughput retroviral tagging to identify components of specific signalling pathways in cancer. Nat Genet 32: 153–159.
27. Lund AH, Turner G, Trubetskoy A, Verhoeven E, Wientjens E, et al. (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat Genet 32: 160–165.
28. Suzuki T, Shen H, Akagi K, Morse HC, Malley JD, et al. (2002) New genes involved in cancer identified by retroviral tagging. Nat Genet 32: 166–174.
29. Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G, Gross F, Yvon E, et al. (2000) Science 288: 669–672.
30. Gaspar HB, Parsley KL, Howe S, King D, Gilmour KC, et al. (2004) Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. Lancet 364: 2181–2187.
31. Wu X, Luke BT, Burgess SM (2006) Redefining the common insertion site. Virology 344: 292–295.
32. Hartung J, Elpelt B, Klösener K-H (1987) Statistik. (Oldenbourg Verlag, München-Wien).
33. Dudewicz EJ, Mishra SN (1988) Modern Mathematical Statistics. (Wiley, New York).
34. Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. Science 300: 1749–1751.
35. Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, et al. (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biology 2: e234.
36. Laufs S, Gentner B, Nagy KZ, Jauch A, Benner A, et al. (2003) Retroviral vector integration occurs in preferred genomic targets in human bone marrow repopulating cells. Blood 101: 2191–2198.
37. Bushman FD (2003) Targeting survival: Integration site selection by retroviruses and LTR-retrotransposons. Cell, 115: 135–138.