Citation: Molecular Therapy — Methods & Clinical Development (2015) **2,** 15015; doi:10.1038/mtm.2015.15

www.nature.com/mtm

ARTICLE

Coherence analysis discriminates between retroviral integration patterns in CD34$^+$ cells transduced under differing clinical trial conditions

Claus V Hallwirth[1], Gagan Garg[1,2], Timothy J Peters[3], Belinda A Kramer[4], Nirav V Malani[5], Jessica Hyman[4], Xiaoan Ruan[6], Samantha L Ginn[1,7], Nicola A Hetherington[1], Lavanya Veeravalli[6], Atif Shahab[6], Shoba Ranganathan[2], Chia-Lin Wei[6], Christopher Liddle[8], Adrian J Thrasher[9], Frederic D Bushman[5], Michael J Buckley[3] and Ian E Alexander[1,10]

Unequivocal demonstration of the therapeutic utility of γ-retroviral vectors for gene therapy applications targeting the hematopoietic system was accompanied by instances of insertional mutagenesis. These events stimulated the ongoing development of putatively safer integrating vector systems and analysis methods to characterize and compare integration site (IS) biosafety profiles. Continuing advances in next-generation sequencing technologies are driving the generation of ever-more complex IS datasets. Available bioinformatic tools to compare such datasets focus on the association of integration sites (ISs) with selected genomic and epigenetic features, and the choice of these features determines the ability to discriminate between datasets. We describe the scalable application of point-process coherence analysis (CA) to compare patterns produced by vector ISs across genomic intervals, uncoupled from association with genomic features. To explore the utility of CA in the context of an unresolved question, we asked whether the differing transduction conditions used in the initial Paris and London SCID-X1 gene therapy trials result in divergent genome-wide integration profiles. We tested a transduction carried out under each condition, and showed that CA could indeed resolve differences in IS distributions. Existence of these differences was confirmed by the application of established methods to compare integration datasets.

Molecular Therapy — Methods & Clinical Development (2015) **2**, 15015; doi:10.1038/mtm.2015.15; published online 29 April 2015

INTRODUCTION

Advances in high-throughput sequencing technologies permit the generation of complex vector integration site (IS) datasets. Comparisons of IS datasets have identified differences in IS distributions attributable to viral origins of vectors used,[1,2] promoter configuration (*e.g.*, wild-type compared to self-inactivating long terminal repeat (LTR) regions),[3] transduced cell types, their gene expression and chromatin conformation,[4] as well as the replicative characteristics of transduced cells.[5] These comparisons have usually been performed by associating integration sites (ISs) with genomic and epigenetic features, where datasets of higher complexity permitted the implementation of increasingly sophisticated analytical comparisons.[2,6]

We sought to develop a method to evaluate vector integration pattern differences, unconstrained by prior knowledge of the genomic or epigenetic contexts of ISs. To this end, we explored the application of binned point-process coherence analysis[7–9] (CA) to detect differences between vector integration patterns. Once IS coordinates have been mapped (a prerequisite shared by all IS analyses), CA facilitates quick and easily implemented pair-wise comparison of IS datasets. A single metric is produced to ascertain the existence of IS pattern differences and thus lends itself to utilization as a proximal screening tool, where the discovery of differences would warrant further investigation into the nature of these differences.

To explore the analytical utility of CA and concurrently begin to address a question of interest to the field of gene therapy, we asked whether this approach is sufficiently sensitive to detect differences in genome-wide γ-retroviral integration patterns in human CD34$^+$ cells induced by the use of different clinical transduction conditions. Accordingly, we applied CA to IS datasets generated from human CD34$^+$ cells transduced under the divergent culture conditions employed in the initial Paris and London SCID-X1 clinical trials. Of the 10 patients enrolled in the Paris trial,[10–12] four developed T-cell acute lymphoblastic leukemia (T-ALL) as a consequence of vector-mediated dysregulation of oncogenes near the site of integration,[13] while only 1 of 10 patients in the London trial developed T-ALL.[14,15]

[1]Gene Therapy Research Unit, Children's Medical Research Institute and The Children's Hospital at Westmead, Westmead, Australia; [2]Department of Chemistry and Biomolecular Sciences, Macquarie University, Macquarie Park, Australia; [3]CSIRO Digital Productivity Flagship, North Ryde, Australia; [4]Children's Cancer Research Unit, The Children's Hospital at Westmead, Westmead, Australia; [5]University of Pennsylvania School of Medicine, Department of Microbiology, Philadelphia, Pennsylvania, USA; [6]Genome Institute of Singapore, Agency for Science, Technology and Research, Genome, Singapore; [7]Sydney Medical School, The University of Sydney, Sydney, Australia; [8]Storr Liver Centre, Westmead Millennium Institute, Sydney Medical School, The University of Sydney, Sydney, Australia; [9]Molecular Immunology Unit, Institute of Child Health, University College London, London, UK; [10]The University of Sydney, Discipline of Paediatrics and Child Health, Westmead, Australia. This work was conducted in Westmead, NSW 2145, Australia. Correspondence: IE Alexander (ian.alexander@health.nsw.gov.au)
Received 23 December 2014; accepted 12 March 2015

Although the difference in the incidence of leukemia is not statistically supported by Fisher's exact test ($P = 0.15$) using such low patient numbers, it raises the question of whether the differing transduction protocols employed could induce distinct integration patterns and, in turn, different leukemogenic risk.

IS datasets were generated through the use of modifications to an established ligation-mediated PCR (LM-PCR) protocol,[16] and by coupling this with Illumina next-generation sequencing technology. CA was used to show that the two samples studied differed, supporting the hypothesis that different transduction conditions produce distinguishable patterns of vector integration. These differences were confirmed by application of established methods to compare vector IS datasets, some of which are consistent with a higher leukemogenic risk associated with the Paris trial transduction conditions.

RESULTS

Differing clinical trial transduction conditions exert phenotypic effects on cultured human CD34$^+$ cell populations

To compare the phenotypic effects of different transduction conditions on transduced cells, human peripheral blood (PB) CD34$^+$ cells were transduced with γ-retroviral vectors according to the transduction protocols employed in the initial SCID-X1 clinical trials conducted in Paris[10,11] and London.[14] Cells cultured under London ("L") trial conditions showed relatively higher retention of CD34 expression at the completion of the transduction period, compared to a loss of CD34-positivity in about half the cells transduced under the Paris ("P") trial conditions (Table 1). Paris conditions, on the other hand, promoted higher levels of both proliferation and transduction. The transduction conditions were shown to exert reproducible effects in terms of CD34$^+$ retention ($P < 0.0001$) and proliferation for a total of two P and three L transductions (Supplementary Table S1), although the difference in proliferation does not reach statistical significance ($P = 0.0553$).

CA distinguishes integration patterns resulting from different transduction conditions

Having observed that the P and L transduction conditions consistently lead to phenotypic differences in transduced cells, we next wished to test whether these populations harbored recognizably different vector integration patterns. IS datasets were generated from P and L cells, yielding 250,215 and 54,424 nonredundant, uniquely mapped ISs, respectively. To begin exploring whether distinguishable patterns are detectable within these IS datasets using point-process CA[9] an initial baseline comparison of independent, nonredundant samples of ISs derived from P was performed (Figure 1a). See Methods, associated references and supplementary material for procedural details of CA. This comparison constitutes a baseline because it is to be expected that patterns characteristic of a particular IS dataset should be retained within subsamples of such a dataset. In accordance with this expectation, this comparison yielded high IS pattern coherence, confirming that the embedded IS patterns within these subsamples are similar. A comparison of P ISs with matched random control (MRC) sites, on the other hand, showed low coherence (Figure 1a), since matched random data would not be expected to contain appreciable patterns. Furthermore, the sensitivity of CA to detect genome-wide IS pattern differences was investigated in a series of semi-quantitative comparisons entailing the deliberate "contamination" of P with random sites. MRC site substitutions of 30, 20, and even 10% could be visually distinguished in terms of coherence (Figure 1a). Next, genome-wide patterns of P and L integration were compared, yielding lower coherence than subsamples of P alone (Figure 1b). This demonstrates the capacity of CA to resolve subtle IS pattern differences in two samples resulting from the use of different transduction conditions. This observation was extended by applying CA to shorter genomic intervals containing lower total numbers of ISs. Consistent with genome-wide CA, P and L pattern differences across a small chromosome could be discerned using a comparatively lower number of ISs associated with this individual chromosome (Figure 1c).

Table 1 Transduction performance under Paris and London SCID-X1 trial conditions

Transduction	*Paris ("P") conditions*	*London ("L") conditions*
Pretransduction CD34$^+$	98.60%	
Final CD34$^+$	49.30%	96.41%
Transgene$^+$	64.85%	16.64%
Transgene$^+$ CD34$^+$	28.81%	15.98%
Proliferation	2.16×	1.74×

Established methods validate integration pattern differences detected by CA

CA was shown to be sufficiently sensitive to detect subtle differences in integration patterns associated with different cell culture conditions employed during transduction in one replicate of each. To investigate the underlying biological nature of these differences, the IS distributions of P and L datasets were compared with respect to selected genomic features. Relative to chance distribution, both datasets displayed the known bias of γ-retroviral vectors[17,18] towards integration in transcription start site (TSS)-proximal regions (Figure 2). The tightness of TSS-clustering, however, was less pronounced in L compared with P (Figure 2a). This finding is also statistically supported by comparisons of the relative proportions of TSS-proximal, intra- and intergenic integration (Figure 2b). In addition to these classifications into three categories relating to coding genes, IS datasets were compared *vis-a-vis* their distributions relative to a selection of annotated genomic features, and in different interval sizes around several of these features (Figure 2c). The direction of the trends towards either over- or underrepresentation at individual features was largely the same for the two transduction datasets. However, the magnitude of the effects of proximity to most annotated features differed markedly between P and L ISs. IS overrepresentation relative to chance distribution was compared in the P and L datasets at all known coding genes, divided into three categories (Figure 3a). Consistent with clustering near TSSs, integration is, on average, overrepresented in all coding genes for both datasets. This is most pronounced at hematological oncogenes and less so at other oncogenes, yet still to a higher degree than at coding genes not associated with neoplasia. When comparing the two datasets, overrepresentation is higher for P than for L integration in all categories, although the difference does not reach statistical significance for hematological oncogenes.

The high complexity of the P and L datasets facilitated statistical comparison of IS counts at individual coding gene loci, using

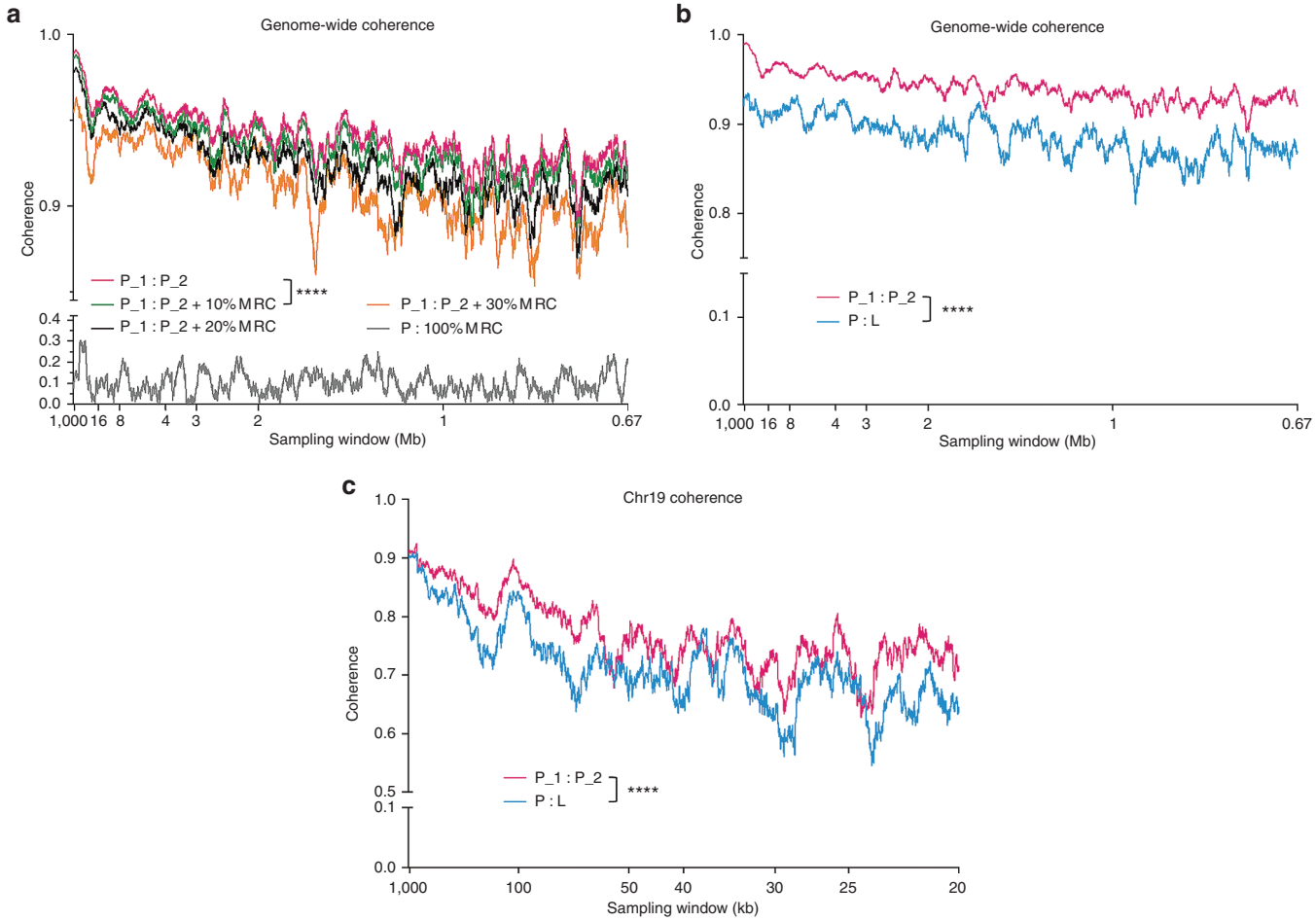

Figure 1 Use of coherence analysis to represent differences in integration site (IS) patterns independently of genomic features. Higher coherence values indicate greater pattern similarity. All dataset subsamples are size-matched for the L dataset (54,424 ISs). (**a**) Coherence values for ISs spanning the entire genome are plotted against an *x*-axis indicating the range of sampling size windows used in the computation of coherence values. Coherence comparisons are shown for two subsamples of P, as well as modifications of one of these subsamples wherein 10%, 20%, 30% or all of the ISs were randomly substituted for matched random control (MRC) sites. (**b**) Coherence values for ISs spanning the entire genome are shown for a comparison of two subsamples of P and for a subsample of P compared to L. (**c**) Coherence values for the same dataset comparisons as panel (**b**), using only those ISs on chromosome 19 (1,607 for the subsample of P and 1,342 for L). ****$P < 0.0001$ for comparison of median coherence values by Wilcoxon matched-pairs signed rank test.

different window sizes around TSSs for counting. This direct comparison of IS counts between the two datasets (as opposed to the consideration of relative overrepresentation in each dataset compared to chance distribution used earlier) identified genes that are differentially targeted by vector integration under each of the two transduction conditions (Figure 3b). The number of genes at which integration is statistically more prevalent under P transduction conditions is considerably larger than that for L conditions. Furthermore, this direct comparison of IS counts at coding genes in P versus L shows that counts across all hematological oncogene loci (taken together) are statistically more prevalent in P than in L, irrespective of the chosen window size around TSSs ($P = 0.00241$ at ±100 kb; $P = 0.01749$ at ±5 kb). Integration is also more prevalent in P compared to L when considering other oncogenes and other coding genes (as gene sets), with statistically greater levels of significance achieved in either window size (data not shown). Wang *et al.*[19] previously reported that the size of chosen intervals around annotated genomic features can affect the magnitude of observed IS clustering. Consistent with this, IS counts at some genes statistically discriminate between P and L only in smaller intervals around the respective TSSs. For instance, IS counts near *LMO2* are statistically greater in P than in L for the TSS ±5 kb interval ($P = 0.0157$), but not for the ±100 kb interval ($P = 0.2755$; Figure 3c).

Comparison of bulk transduced CD34⁺ cell- and patient-derived IS patterns

To investigate integration pattern differences of transduced cells *in vitro* and *in vivo*, IS data from PB mononuclear cells (PBMCs) recovered from patients treated in the Paris and London SCID-X1 trials[19,20] (hereafter referred to as "SCID1_Paris" and "SCID1_London") were compared to one another and with the P and L data. The two patient datasets display more coherent patterns than the comparison between P and L (Figure 4a,b); the coherence values are even higher, in fact, than for intra-dataset comparisons using subsamples of either the P or the L dataset (Figure 4b). Wang *et al.*[19] previously reported that less than 1% of cells infused in the Paris trial had long-term repopulating potential. Such cells would presumably have given rise to the PBMC IS datasets, and be distinguished from the bulk transduced CD34⁺ cells by displaying a distinctive immunophenotype. The greater pattern coherence between IS datasets derived from the two different clinical trials implies that the immunophenotypic

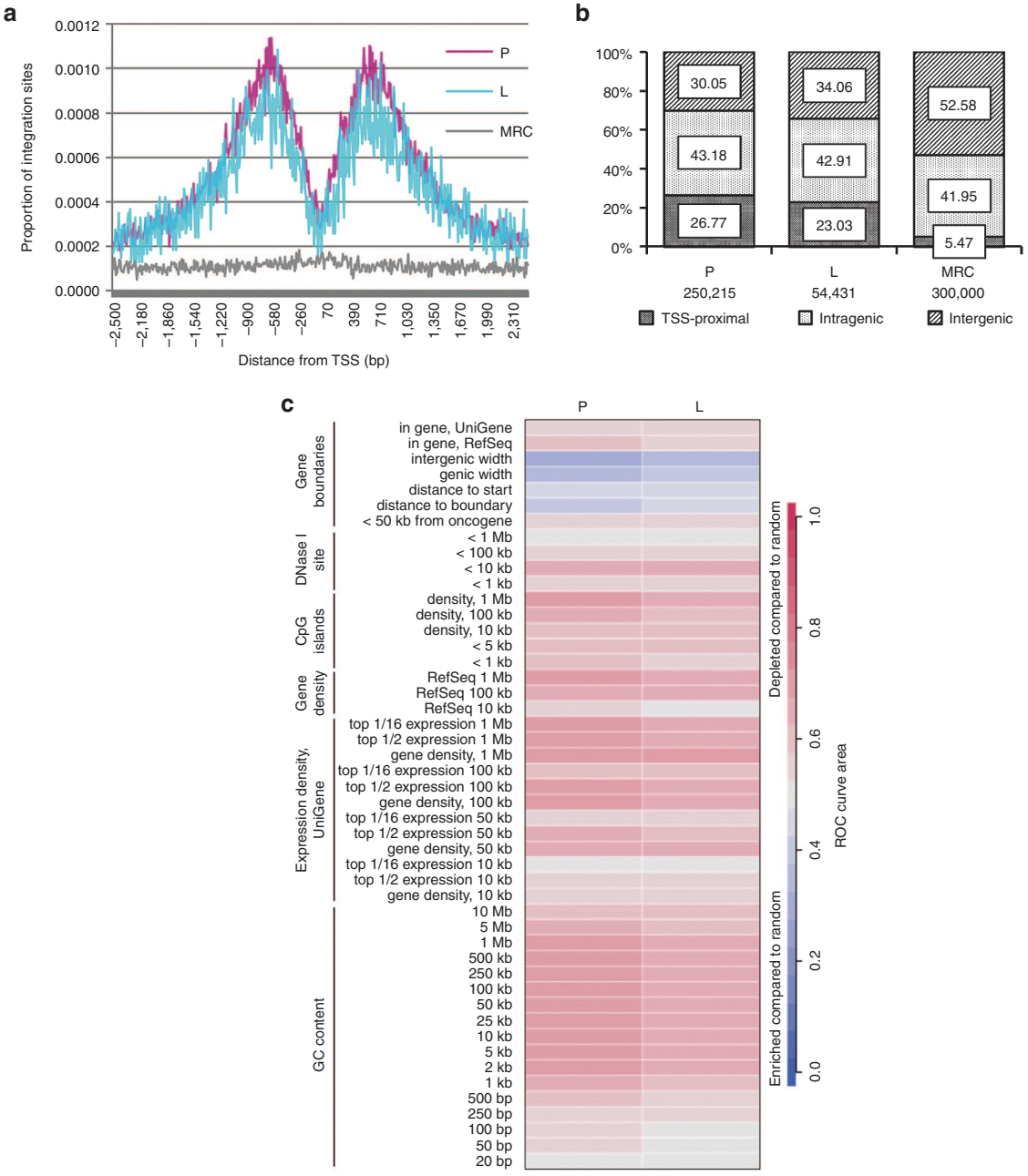

Figure 2 Distribution of γ-retroviral vector integration sites (ISs) relative to genomic features. (**a**) Distances from ISs to the nearest "UCSC Known Gene" transcription start site (TSS) were computed, binned in 10-bp windows, converted to proportions relative to the total number of sites in the respective datasets, and are plotted for those sites falling within 2.5 kb either side of the nearest TSS. MRCs, matched random control sites. (**b**) Integration site distributions relative to coding genes. Percentages of sites within each genic category are shown in the columns, and the total number of sites per dataset is indicated below. Comparison of IS distributions between P and L datasets shows very strong statistical evidence of differences ($\chi^2 = 473$, df = 2, $P < 0.0001$). The proportions of intragenic integration do not differ between the two datasets (two-tailed Fisher's exact test on IS counts, $P = 0.26$). The *P* values for all comparisons within genic categories are presented in Supplementary Tables S3–S5. (**c**) The extent of association between annotated genomic features and vector ISs is summarized in the form of heat maps. Columns are labeled with IS dataset names and rows with analyzed genomic features. The colored receiver operating characteristic (ROC) curve area scale to the right of the panel shows increased integration (relative to MRC sites) near the indicated feature in red, decreased integration in blue, with the intensity of shading correlating with the degree of departure from random integration. Detailed comparisons of relative integration abundance associated with individual genomic features are available as "Supplementary report - Association of Genomic Features with Integration".

subset of CD34⁺ cells being transduced has a greater influence on the resulting integration patterns than the differences imparted by use of P versus L transduction conditions. Consistent with this, the intra-dataset comparison of SCID1_Paris ISs exhibits the highest coherence amongst all the comparisons conducted (Figure 4b). Inter-dataset coherence between patient datasets and either P or L is lower than the comparison of P and L. Amongst these latter comparisons, both P and L patterns are more coherent with those of SCID1_London than SCID1_Paris (Figure 4b; $P < 0.0001$).

Statistical analysis of relative proportions of TSS-proximal, intra- and intergenic integration (Supplementary Table S2) revealed intragenic integrations to be more abundant in the two CD34⁺ cell datasets

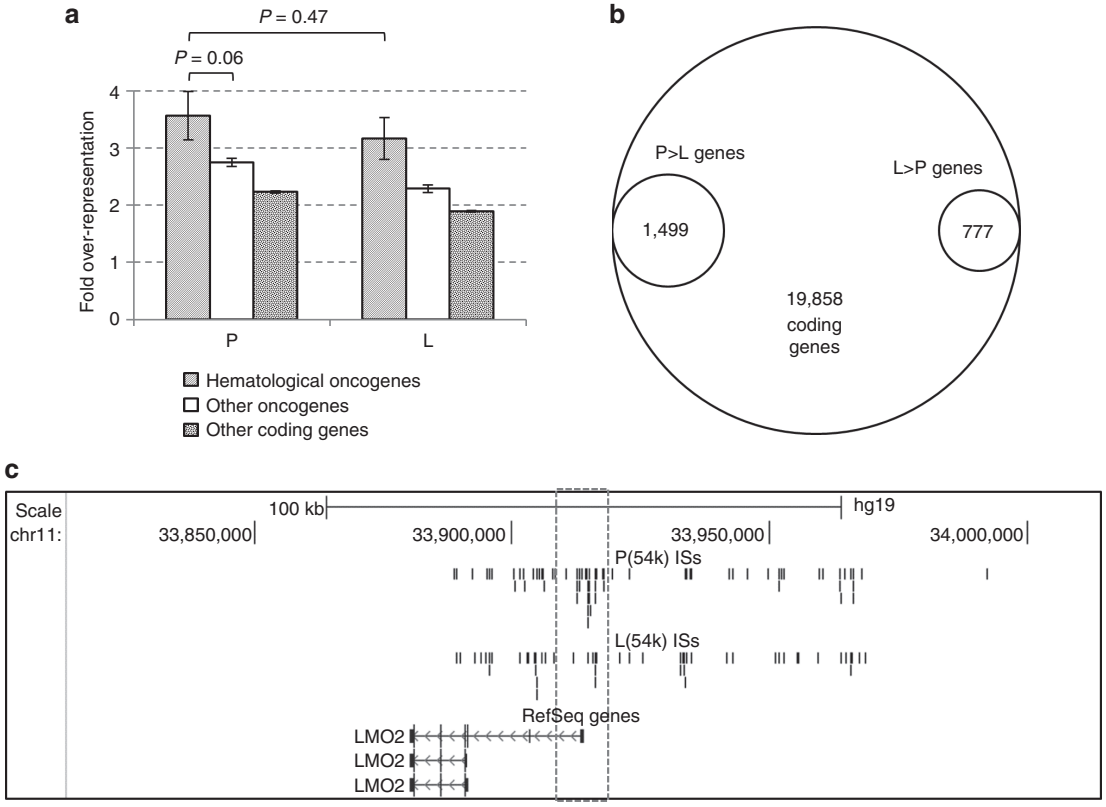

Figure 3 Over-representation of γ-retroviral vector integration sites (ISs) at coding genes. (**a**) Overrepresentation values relative to random sites were calculated for proportions of ISs falling within ±100 kb of the transcription start site (TSS) of each known coding gene. A subset of oncogenes was extracted from this list of genes, leaving "other coding genes" (n = 17,822). Oncogenes associated with hematological malignancies were designated "hematological oncogenes" (n = 89), leaving "other oncogenes" (n = 1,852). All comparisons of mean overrepresentation values showed statistical support of differences (independent t-tests, $P < 0.05$), except where indicated. Error bars indicate the standard errors of the means. (**b**) Genes exhibiting higher integration frequency under either P or L transduction conditions. Numbers of genes where IS counts within TSS ±100 kb in one dataset statistically outnumber counts in the other dataset (one-sided Fisher's exact test, $P < 0.05$) are shown in the internal circles. (**c**) Coordinates of P(54k) (size-matched for L) and L ISs falling within 100 kb of the *LMO2* TSS are shown (http://genome.ucsc.edu). The region within 5 kb of the *LMO2* TSS is boxed.

compared to ISs recovered from patient PBMCs. When IS distributions are compared at selected genomic features, the PBMC datasets, on average, display greater deviation from random integration than the CD34$^+$ cell datasets (Supplementary Figure S1). Similarly, integrations are more highly overrepresented in SCID1_Paris than in either P or L within 100 kb of the TSS of hematological oncogenes and other oncogenes, and to a lesser extent, other coding genes (Supplementary Figure S2). The SCID1_London dataset is not sufficiently complex to be included in this analysis. When proportions of ISs at individual hematological oncogenes are directly compared to one another (as opposed to being expressed as overrepresentation relative to random sites), all IS datasets display a wider range of proportional integration than MRCs, indicating that some hematological oncogene loci are more prone to vector integration than others (Figure 5). Amongst the IS datasets, SCID1_Paris exhibits higher proportional integration at most hematological oncogenes, while IS proportions in P and L are nearly equivalent. Again, the SCID1_London dataset is not sufficiently complex to be included in this analysis. In summary, the measurable behavior of ISs relative to coding genes and other annotated genomic features are more pronounced in PBMCs recovered from patients than in bulk populations of transduced CD34$^+$ cells.

DISCUSSION

The generation of high-complexity datasets to study viral integration behavior is becoming increasingly more attainable, and as such, the need for quick and simple comparisons of such datasets is likely to increase. To this end, we transitioned the use of CA, rooted in the work done on time series and point processes by Brillinger in the 1970s,[7] from neuroinformatics into the genomic paradigm, and introduce CA as a tool to compare genomic IS distribution patterns resulting from the transduction of cells with integrating vectors. Because CA directly compares patterns, it requires no prior knowledge of the association of ISs with genomic or epigenetic features. Established methodologies are sometimes constrained, to a degree, by the need to conduct numerous sequential analytical comparisons on a considerable diversity of features, or a range of intervals surrounding static genomic features, before identifying individual parameters that distinguish IS datasets. CA, on the other hand, produces single metrics that can be visually interpreted and their differences statistically validated by comparison of median coherence values. For this reason, and because datasets are directly compared without the often time-consuming generation of MRCs (as is required to produce heat maps), CA is more economical of both operator and computing time and thus lends itself to utilization as a proximal screening tool, where the detection of differences would warrant further investigation to elucidate the nature and origins of such differences, using more specific methods.

CA can be scaled to accommodate published (often smaller) as well as future, larger datasets. Although the lower limit of usable dataset sizes was not empirically determined as part of this study,

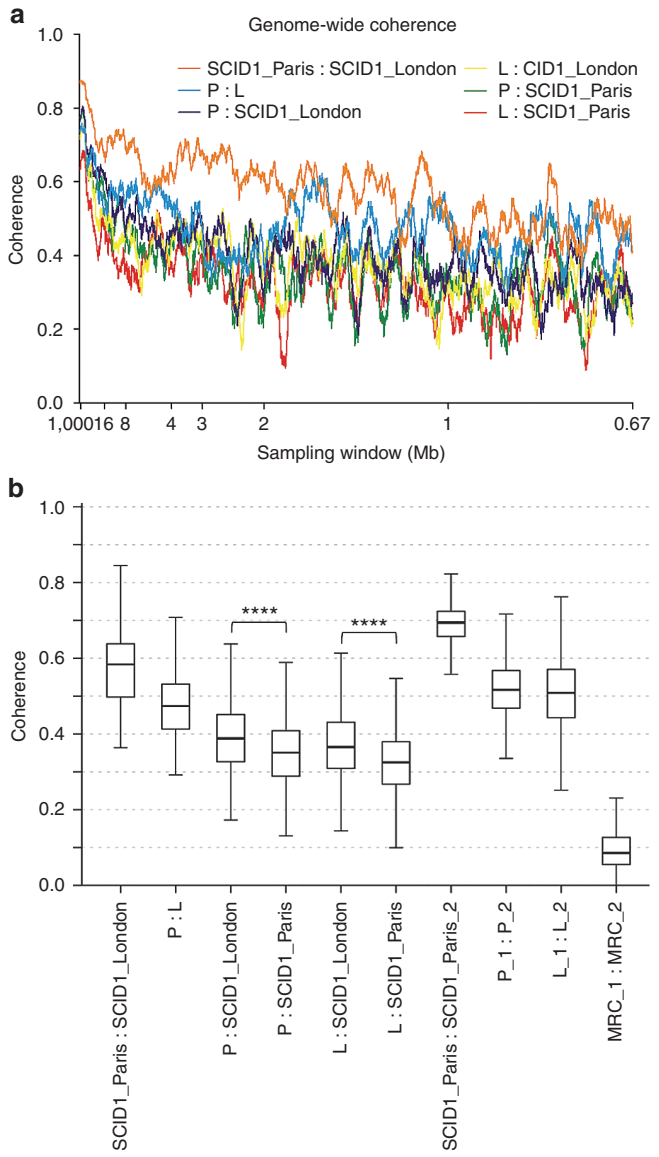

Figure 4 Coherence analysis comparisons of integration site (IS) patterns in transduced CD34$^+$ cells and peripheral blood mononuclear cells from patients treated for SCID-X1. Integration site datasets were randomly subsampled to match the size of the SCID1_London dataset (3,870 ISs). (**a**) Whole-genome coherence values are plotted for comparisons of IS patterns between the P, L, SCID1_Paris and SCID1_London datasets. (**b**) The distributions of coherence values representing the individually plotted values in panel (**a**) and additional intra-dataset comparisons are represented as Tukey box plots:[36] lower and upper bounds of boxes depict the first and third quartiles, respectively; the horizontal lines in the boxes indicate the median coherence values; bottom and top whiskers represent values corresponding to 1.5× the interquartile range of the lower and upper quartiles, respectively; and outliers and extreme values are not shown. Whole-genome comparisons comprise 8,191 data points. ****$P < 0.0001$ for comparison of median coherence values by Wilcoxon matched-pairs signed rank test (only shown for two selected statistical comparisons referred to in Results).

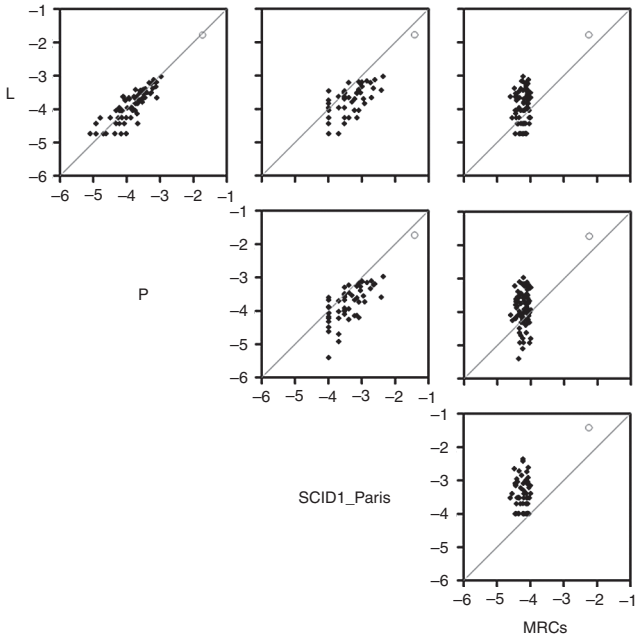

Figure 5 Integration at hematological oncogenes. Proportions of integration sites in P, L, SCID1_Paris and matched random controls (MRCs) located within 100 kb of the transcription start sites of genes associated with hematological malignancies are plotted relative to one another by dataset. Individual hematological oncogenes are represented by closed diamonds. The open circle on each plot represents the proportion of counts across all these genes, relative to the size of the respective dataset. The diagonal line in each plot represents equal proportions, with genes falling on the side of higher proportional representation. The axes are labeled with the exponents of $\log_{10}$ proportions. Both over- and under-representation relative to MRCs at individual loci can be calculated for P and L (right column), while only overrepresentation can be ascertained at individual loci for SCID1_Paris, wherein even a single IS within a 200-kb interval constitutes overrepresentation relative to MRCs (bottom right panel).

size-matched datasets of 3,870 ISs distributed across the whole genome were compared to yield statistically validated observations (Figure 4). In a smaller genomic interval, such as a single chromosome, fewer than 1,700 ISs sufficed to confirm pattern differences observed at the genome-wide scale. It is likely that the magnitude of pattern differences would affect the minimum number of ISs required in order to detect such differences. CA accommodates the comparison of differently sized datasets and corrects for size differences. We did observe, however, that the use of larger IS samples from our datasets yielded higher coherence values (data not shown). More complex datasets represent a larger sample of an IS "population", and are thus more accurate in resolving the true underlying IS pattern in any given sample. The sensitivity of CA was established by demonstrating that the contamination of an IS dataset with as few as 10% random sites was detectable. Although contamination with 30% random sites yields coherence values that are higher than those obtained when comparing Paris to London transduction conditions (*cf.* Figure 1a,b), it is to be borne in mind that differences between independently generated IS datasets are likely to be systematic rather than random and it would likely be incorrect to infer that patterns resulting from Paris versus London transduction conditions differ by >30%. Pattern deviations resulting from contribution of random sites are unsuitable as measures of calibration to gauge the extent of observed pattern differences between datasets.

As an exemplar of the potential analytical utility of CA, it was shown to be sufficiently sensitive to distinguish differences between IS patterns in a pair of samples differing only by the transduction conditions used. Specifically, phenotypic differences were noted between cells transduced under either the Paris or London SCID-X1 clinical trial conditions. The phenotypic characteristics were

consistent between three independent transductions performed under London conditions (Supplementary Table S1), whereas high proliferation and loss of CD34-positivity in cells transduced under Paris conditions were consistent with cells transduced under the same conditions as part of our treatment of a SCID-X1 patient.[12] Essentially, Paris conditions likely led to greater immunophenotypic heterogeneity at the end of the transduction protocol. The transduction conditions employed in this study differed principally with regard to the absence or presence of serum during vector production and transduction, and the concentration of IL-3. The two trials also used packaging cell lines expressing different envelope proteins. This confounding factor was absent in the current study, where all vectors were GALV-pseudotyped.

We explored a lingering question pertaining to possible consequences of the different transduction conditions employed in the Paris and London SCID-X1 trials by determining whether the phenotypic differences imparted by these conditions are recapitulated at the level of vector IS distributions. In recognition of the fact that junction fragment library construction was not replicated, consistency of IS pattern coherence was first established within each of the IS datasets generated. The primary aim was to ascertain the capacity of CA to detect potentially very small pattern differences between independent IS datasets. In fulfillment thereof, CA was used to detect differences in IS patterns in cells transduced with the same vectors, but using either the Paris or London culture conditions during transduction. Such differences are expected to be more subtle than those attributable to the use of γ-retro- versus lentiviral vectors, or transduction of different cell types, although the use of the two transduction conditions in this study did lead to different immunophenotypic distributions. We assayed only a single sample from each condition, so within-condition variation is uncharacterized and represents an alternative explanation for the difference detected. CA revealed that integration patterns are most coherent amongst PBMCs recovered from patients treated in the Paris and London SCID-X1 trials, as opposed to patterns from transduced $CD34^+$ cells being compared to these clinical datasets. This observation is likely attributable to the fact that bulk transduced $CD34^+$ cells characterized by heterogeneous immunophenotypes were compared to cells that arose from the expansion of transduced true hematopoietic progenitor cells (HPCs) characterized by narrowly defined immunophenotypes. Furthermore, integration events in bulk transduced cells are recoverable irrespective of whether or not they are located favorably for transgene expression, whereas only cells whose progenitors harbored productive integrations are expected among PBMCs of treated patients. Furthermore, both bulk transduced integration patterns were more coherent with PBMC patterns from London patients than with Paris patients (Figure 4b). This could conceivably be attributed to the use of an amphotropic envelope in the Paris trial, compared to a GALV envelope in the London trial as well as both of our $CD34^+$ cell transductions.

The integrity of the CA findings was confirmed by investigating the specific underlying differences in integration behavior in the two samples transduced under the Paris and London transduction conditions. This entailed the use of established analytical methodologies to compare IS datasets, relying upon associations of IS distributions with selected genomic features. Implementation of methodological strategies to improve recovery of junction fragments led to datasets of such high complexity that they facilitated statistical comparison of IS counts at every known coding gene in the genome. These comparisons revealed that up to 7.5% of all coding genes display a statistically higher frequency of integration in one transduction dataset compared to another, though again we note that the variation among replicates for a single condition has not been studied. The number of discriminatory genes is considerably larger for the Paris transduction conditions, most probably as a consequence of the greater phenotypic diversity among this transduced cell population. The oncogene *LMO2*, dysregulation of which was linked to the development of leukemia in four of the five patients in the SCID-X1 trials, was amongst the loci displaying statistically higher levels of integration under Paris compared to London transduction conditions, but only in a narrow interval around the TSS. This is consonant with the tighter IS clustering around TSSs resulting from the Paris conditions, when averaged for all genes. It is noteworthy that two of the three clonal *LMO2* integrations amongst the patients that developed leukemia in the Paris SCID-X1 trial were within 5 kb of the *LMO2* TSS (the third was at a distance of 10.6 kb), whereas the *LMO2* integration in the London trial leukemia case was 35 kb from the TSS. Direct comparison of integration counts revealed a statistically supported bias in favor of Paris over London transduction conditions near hematological oncogenes. The level of significance achieved for oncogenes not specifically associated with hematological malignancies (collectively, as a gene set) was much greater than for hematological oncogenes, given the considerably larger numbers of loci comprising this gene set. Again, this is consistent with the tighter TSS-proximal IS clustering under Paris conditions. Overall, the observed dissimilarities would support a hypothesis of Paris transduction conditions imparting a higher oncogenic risk than the London conditions, consistent with the observed incidence of leukemia in the two trials. However, definitive conclusions will require analysis of multiple replicates under each condition and the characterization of integration profiles amongst subpopulations of transduced cells specifically representing uncommitted HPCs.

The pronounced clustering effects resulting from Paris compared to London transduction conditions are despite the greater phenotypic heterogeneity amongst the former transduced cell population. Such phenotypic variance would dilute the observable integration patterns, since phenotypically different cells are likely to exhibit different IS patterns.[4] This suggests that the actual effects of the Paris trial transduction protocol on true HPCs might be more accentuated than the observable effects in a bulk population of transduced cells. This is supported by the even more pronounced effects seen in the Paris patient PBMC dataset (Figure 5 and Supplementary Figure S1). The PBMCs from which this dataset was derived would have originated from a narrowly defined set of HPCs, and the differences seen between bulk $CD34^+$ cells transduced under Paris conditions and PBMCs from Paris patients are probably accounted for by phenotypic heterogeneity and homogeneity, respectively. An immediately apparent interpretation of the higher proportion of integration at hematological oncogenes in patient-derived cells compared to bulk transduced $CD34^+$ cells would be that all such integration events individually imparted a survival advantage, consistent with nonmalignant clonal expansions of transduced cells.[21] Alternatively, the integration profile observed in patient cells could reflect the specific subtype of $CD34^+$ cells, at the time of transduction, with long-term repopulating potential. More of the genes involved in hematopoietic development and regulation would be active and accessible to integration in these cells than in a bulk population of $CD34^+$ cells, only a small proportion of which represents true HPCs. These two explanations could be conceptually linked if a pronounced loss of true HPCs under Paris conditions leads to increased proliferative pressure on the remaining HPCs, resulting in selection of clones with integrations near hematological

oncogenes. A further comparison of ISs in transduced CD34$^+$ cells versus PBMCs recovered from treated patients yields potential insight into the impact of vector integration on gene function at different stages of cellular differentiation. Intragenic integrations were more abundant in both our bulk transduced cell datasets than in patient PBMCs (Supplementary Table S2). This contrasts with the reported lack of difference in intragenic integration frequency in transduced PB T-cells pre- and post-transfusion,[22] yet is in agreement with higher abundance of intragenic integration in CD34$^+$ cells pre-gene therapy compared to CD3$^+$ and CD15$^+$ cells post-gene therapy.[4] These apparently discrepant findings could be reconciled by consideration of the following postulate: A larger proportion of progenitor cells is rendered nonviable by integrations within transcriptional units of genes required during hematopoietic development and differentiation, compared to the proportion of already differentiated T-cells harboring integrations in such genes.

In conclusion, we introduce CA as a tool that can sensitively detect subtle differences in vector IS patterns, uncoupled from their association with specific genomic or epigenetic features. We showed that CA can distinguish IS patterns induced by the different cell culture conditions employed during transduction as part of the Paris and London SCID-X1 trials in one replicate of each. These pattern differences were validated by application of established methodologies to compare IS datasets, which warrant further investigation of the hypothesis that Paris transduction conditions could engender IS distributions carrying a higher risk of oncogenesis than the London trial conditions.

MATERIALS AND METHODS

Cells and transduction conditions

Hematopoietic stem cells mobilized in a 7-year-old male pediatric oncology patient using G-CSF and chemotherapy were CD34-selected using an Isolex 300i device (Baxter, Old Toongabbie, Australia) and cryopreserved. All cell stocks used in this study were no longer required for clinical application, and available for research under a Human Research Ethics approval at the Children's Hospital at Westmead. CD34$^+$ cells were transduced with independent batches of MGMT-encoding MFG-based γ-retroviral vectors collected from the supernatant of PG13 producer cells. Vector stocks for the London conditions were serum-free, containing 1% human serum albumin (Albumex 20, CSL, Parkville, Australia). Vector stocks for the Paris conditions contained 4% FCS (v/v, Invitrogen, Mulgrave, Australia). One transduction was carried out under conditions employed in the Paris SCID-X1 trial[10,11] and three under the London SCID-X1 trial[14] conditions (Table 1 and Supplementary Table S1). Detailed transduction protocols are provided as Supplementary Methods. It is to be noted that this study did not attempt to copy the clinical trial conditions in their entirety (*e.g.*, cell selection procedures and transgenes differed from the trials), such that the results do not translate directly to the clinical trials. A one-tailed unpaired *t*-test was used to compare fold-proliferation values (Supplementary Table S1). Percentage retention of CD34-positivity (not final CD34-positivity; Supplementary Table S1) was compared using a two-tailed *t*-test. The CD34$^+$ retention of the L condition replicate wherein CD34-positivity increased during the transduction protocol was taken as 100%.

Junction fragment library construction

Genomic DNA from transduced cells was extracted using a Puregene Blood and Cell Culture DNA Kit (Qiagen, Chadstone Centre, Australia), according to the manufacturer's protocol for cultured cells. DNA, eluted in DNA Hydration Solution, was stored at −20 °C until use. A previously described LM-PCR method[16] was employed to selectively amplify junction fragments comprising LTR-derived proviral DNA and adjoining host DNA sequences, as detailed in Supplementary Methods. Apart from primer design specific to the use of an Illumina NGS platform, the method was adapted to improve linker ligation efficiency via partial filling of 5′ overhangs after gDNA digestion with *Tsp*509I (New England Biolabs, Genesearch, Arundel, Australia) and the ligation of adapters compatible with the partially filled ends. This approach prevents religation of restricted gDNA. Furthermore, to accommodate input fragment length limitations pertaining to library sequencing on the Illumina Genome Analyzer IIx (GAIIx) platform, LM-PCR amplicons >400 bp were reprocessed using two additional restriction endonucleases (REs) with four-base recognition sequences: *Mbo*I (New England Biolabs) and *Csp*6I (Thermo Fisher Scientific, Scoresby, Australia). This ensured that LM-PCR amplicons that would otherwise have been too large to facilitate efficient bridge amplification could be sequenced, thus increasing the recovery of ISs. While the choice of the RE for methods utilizing RE digestion is known to bias IS recovery,[23,24] the ascertainment biases apply equally to both junction fragment libraries in this study.

Fragment library sequencing and mapping of ISs

Synchronous sequence homogeneity in the first 32 positions of all reads (arising from the vector LTR) was accommodated by applying the spectral calibration parameters of an adjacent flow cell lane containing balanced heterogeneous nucleotide distributions from another library, thereby facilitating correct base calling. Relevant reads from P and L libraries were identified using Bowtie[25] by searching for the 28-base LTR sequence contributed to the amplicons by the primer MLVN1 (Supplementary Methods), allowing two mismatches or 5′ truncations. Reads were also required to contain the sequence TTCA in positions 29–32, derived from amplification primed specifically off the LTR. A custom Perl script was used to trim MLVN1- and LTR-derived proximal sequences. Potential distal adapter sequences were identified using Bowtie and trimmed using another custom Perl script. The approach was to screen reads for distal adapter sequences of 26 bases; the remainder for 25 bases; and iteratively down to only a single base (G). The search allowed up to two mismatches for distal adapter sequences of 26-20 bases, one mismatch for 19-10 bases, and no mismatch below 10 bases. A further condition for recognition of putative distal adapter sequences was that they were preceded by the recognition sequences of *Tsp*509I, *Mbo*I, or *Csp*6I. After trimming of distal adapter sequences, reads were pooled with those not found to contain any sequence contributions from the distal adapter. The reads ranged in length from 44 bases to 18 bases (if the maximum 26-base distal adapter contribution was trimmed). Reads ≥20 bases were mapped to human reference sequences hg19/GRCh37 and to hg18/NCBI36 using Bowtie, allowing no mismatches and disregarding any reads that mapped to more than one location under these conditions. Reads mapped with zero mismatches were mapped again, this time allowing two mismatches. Any reads that mapped to more than one location under these conditions were disregarded. This strategy was implemented to reduce false-positive mapping of any reads having been generated on the GAIIx platform with one or two base position errors. The output from Bowtie was set to SAM format,[26] which was further translated into BED format using BEDtools[27] to obtain the coordinates of ISs in the human genome. The base pair prior to the first position of every mapped forward orientation read was taken to denote the position of the respective integration event. Similarly, the last base pair of every reverse orientation read was taken as the position of integration.

Generation of matched random control sites

A set of MRC sites was computationally generated for statistical comparison with the different IS datasets. These MRCs were selected based on the use of the *Tsp*509I RE during junction fragment library generation, and the size selection of the LM-PCR amplicons. All genomic positions greater than 18 bp and less than 360 bp from *Tsp*509I sites were identified using a custom Perl script, yielding 1,453,022,637 and 1,454,239,958 unique locations for hg18 and hg19, respectively, wherein true integration events could theoretically be detected by the junction fragment recovery method employed. MRC site sets of different sizes were obtained using the random number generator function in Perl to suit individual analysis needs, as used in similar applications previously.[2,28] After identification, randomly selected positions were mapped back to chromosomes in proportions reflective of the relative chromosome sizes.

Coherence analysis

CA assesses the association between a pair of point processes, and does so in the frequency (Fourier) domain. In our case, the point processes are collections of ISs on the human genome, and frequency *f* can be thought of as a reciprocal of genomic window size *w*:

$$f \sim 1/w \quad (1)$$

The coherence C_{xy} of two point processes x and y, for a given frequency f, is defined as

$$C_{xy}(f) = \frac{\left|G_{xy}(f)\right|^2}{G_{xx}(f).G_{yy}(f)} \quad (2)$$

where $G_{xx}(f)$ and $G_{yy}(f)$ are the spectral densities of the two processes and $G_{xy}(f)$ is the cross-spectral density. The spectral density of a point process is the Fourier transform of its autocovariance density, while the cross-spectral density of two point processes is the Fourier transform of the cross-covariance density.[29] $C_{xy}(f)$ can be interpreted as the degree to which the two point processes are in phase with each other at frequency f. For every IS dataset included in CA, IS coordinates across all chromosomes were concatenated into a single-vector sequence, essentially converting the human genome into a single chromosome for computational purposes. Coherence values were computed using chronux (http://chronux.org/),[30] an open-source software package developed for analysis of neurobiological time series data implemented in MATLAB (MATLAB, Statistics and Signal Processing Toolbox Release 2012b, MathWorks, Sydney, Australia). All single-vector IS coordinates were rescaled by a divide factor d, proportional to the full range of coordinates being considered for any genomic interval tested (parameter: delay_times; see Supplementary Table S6). This is necessary to ensure computational feasibility when dealing with large numerical ranges of point processes, in this instance coordinates pertaining to the human genome. After the coherence values were plotted across the frequency domain F, the corresponding nucleotide window sizes (Figure 1, x-axis labels) w were obtained using $w = d/F$. After ascertaining that the chronux coherencypt() function produces the same comparative results as the coherencypb() function for subgenomic intervals ≤1 Mb, coherencypt() was used for comparative analyses of larger intervals, owing to the prohibitively high memory utilization of coherencypb() in the context of our particular application. Coherencypt() (parameters: Fs, fpass, err, Pad and tapers) was used to compute coherence values for all pair-wise comparisons of IS datasets. CA was performed across a range of genomic interval sizes (Figure 1; see Supplementary Table S6 for details of parameter choices), with analysis of shorter genomic intervals computationally facilitating the sampling of datasets down to 20-kb windows. All subsamples of IS datasets were randomly sampled without replacement to avoid identical IS coordinates in different samples. Median coherence values were compared using Wilcoxon matched-pairs signed rank test, where the pairing of data refers to matched coherence values for the same frequency values.

Analysis and annotation of γ-retroviral ISs

Individual ISs were mapped to the closest University of California Santa Cruz (UCSC) known genes[31] (UCSC Known Genes database: February 2012) TSSs using BEDtools. IS frequencies were counted in window sizes of 10 bp. For all genomic feature mapping applications, numbers of MRCs suitable for use as controls were randomly selected. ISs of all datasets were also classified as TSS-proximal, intragenic and intergenic relative to UCSC known genes, defined as follows: (i) transcription start site (TSS)-proximal when located within 2.5 kb upstream or downstream of a TSS; (ii) intragenic when located between the TSS and transcription end site (except for those sites that fulfil the criterion of being TSS-proximal); and (iii) all remaining sites are regarded as intergenic.[2] For the generation of IS heat maps relative to annotated genomic features, over- and underrepresentation was calculated by statistical comparison against MRC sites using the receiver operating characteristic (ROC) curve area method.[6,32]

Statistical analysis of IS counts across defined gene intervals

ISs were counted in intervals of TSS ±100 kb (the distance at which strong LTR-encoded enhancer/promoter elements can still upregulate gene expression bidirectionally[33]) and ±5 kb at 19,885 coding gene loci across the human genome. All coding genes were divided into three discrete categories by extracting a subset of oncogenes, leaving "other coding genes" (n = 17,822), and from within the oncogenes, those associated with hematological malignancies ("hematological oncogenes", n = 89), leaving "other oncogenes" (n = 1,852). Oncogenes were defined as all coding genes in the "allOnco" list at http://www.bushmanlab.org/links/genelists. Hematological oncogenes comprise nonredundant genes, combining lists "humanLymph" from the same URL and the "Leukemia Gene Database" (www.bioinformatics.org). Interval counts were expressed as proportions of the total number of sites in the respective datasets. Over-representation ratios were computed based on IS proportions divided by MRC proportions for the same intervals. This calculation cannot be done where the MRC count is zero (as was the case at one locus) and is unstable where the MRC count is less than 20. For these reasons, 122 genes were omitted from this analysis. Mean over-representation scores pertaining to counts in the TSS ±100 kb intervals for the three gene categories were compared within and between datasets, using independent t-tests (Figure 3a). To identify genes at which integration frequencies statistically discriminate between the P and L datasets, one-sided Fisher's exact tests for 2 × 2 contingency tables[34] were performed using the IS counts in TSS ±100 kb and ±5 kb intervals of all coding genes. ISs near *LMO2* were plotted using the UCSC Genome Browser.[35]

CONFLICT OF INTEREST

The authors declare no conflict of interest

ACKNOWLEDGMENTS

Funding to support for this work was provided in part by an International Science Linkages grant (CG130052) from the Australian Department of Innovation, Industry, Science and Research and a project grant (APP1026710) from the Australian National Health and Medical Research Council.

REFERENCES

1. Mitchell, RS, Beitzel, BF, Schroder, AR, Shinn, P, Chen, H, Berry, CC *et al.* (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**: E234.
2. Cattoglio, C, Pellin, D, Rizzi, E, Maruggi, G, Corti, G, Miselli, F *et al.* (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* **116**: 5507–5517.
3. Cavazza, A, Cocchiarella, F, Bartholomae, C, Schmidt, M, Pincelli, C, Larcher, F *et al.* (2013). Self-inactivating MLV vectors have a reduced genotoxic profile in human epidermal keratinocytes. *Gene Ther* **20**: 949–957.
4. Biasco, L, Ambrosi, A, Pellin, D, Bartholomae, C, Brigida, I, Roncarolo, MG *et al.* (2011). Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol Med* **3**: 89–101.
5. Bartholomae, CC, Arens, A, Balaggan, KS, Yáñez-Muñoz, RJ, Montini, E, Howe, SJ *et al.* (2011). Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol Ther* **19**: 703–710.
6. Berry, C, Hannenhalli, S, Leipzig, J and Bushman, FD (2006). Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol* **2**: e157.
7. Brillinger, DR (1978). Comparative aspects of the study of ordinary time series and of point process. *Developments in Statistics*. Vol. 1. Academic Press, New York. pp. 33–133.
8. Jarvis, MR and Mitra, PP (2001). Sampling properties of the spectrum and coherency of sequences of action potentials. *Neural Comput* **13**: 717–749.
9. Brown, EN, Kass, RE and Mitra, PP (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nat Neurosci* **7**: 456–461.
10. Cavazzana-Calvo, M, Hacein-Bey, S, de Saint Basile, G, Gross, F, Yvon, E, Nusbaum, P *et al.* (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**: 669–672.
11. Hacein-Bey-Abina, S, Le Deist, F, Carlier, F, Bouneaud, C, Hue, C, De Villartay, JP *et al.* (2002). Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med* **346**: 1185–1193.
12. Ginn, SL, Curtin, JA, Kramer, B, Smyth, CM, Wong, M, Kakakios, A *et al.* (2005). Treatment of an infant with X-linked severe combined immunodeficiency (SCID-X1) by gene therapy in Australia. *Med J Aust* **182**: 458–463.
13. Hacein-Bey-Abina, S, Garrigue, A, Wang, GP, Soulier, J, Lim, A, Morillon, E *et al.* (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* **118**: 3132–3142.
14. Gaspar, HB, Parsley, KL, Howe, S, King, D, Gilmour, KC, Sinclair, J *et al.* (2004). Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet* **364**: 2181–2187.
15. Howe, SJ, Mansour, MR, Schwarzwaelder, K, Bartholomae, C, Hubank, M, Kempski, H *et al.* (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J Clin Invest* **118**: 3143–3150.
16. Ciuffi, A, Ronen, K, Brady, T, Malani, N, Wang, G, Berry, CC *et al.* (2009). Methods for integration site distribution analyses in animal cell genomes. *Methods* **47**: 261–268.
17. Roth, SL, Malani, N and Bushman, FD (2011). Gammaretroviral integration into nucleosomal target DNA in vivo. *J Virol* **85**: 7393–7401.

18. Wu, X, Li, Y, Crise, B and Burgess, SM (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**: 1749–1751.
19. Wang, GP, Berry, CC, Malani, N, Leboulch, P, Fischer, A, Hacein-Bey-Abina, S *et al.* (2010). Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood* **115**: 4356–4366.
20. Schwarzwaelder, K, Howe, SJ, Schmidt, M, Brugman, MH, Deichmann, A, Glimm, H *et al.* (2007). Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *J Clin Invest* **117**: 2241–2249.
21. Fehse, B and Roeder, I (2008). Insertional mutagenesis and clonal dominance: biological and statistical considerations. *Gene Ther* **15**: 143–153.
22. Cattoglio, C, Maruggi, G, Bartholomae, C, Malani, N, Pellin, D, Cocchiarella, F *et al.* (2010). High-definition mapping of retroviral integration sites defines the fate of allogeneic T cells after donor lymphocyte infusion. *PLoS One* **5**: e15688.
23. Wang, GP, Garrigue, A, Ciuffi, A, Ronen, K, Leipzig, J, Berry, C *et al.* (2008). DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res* **36**: e49.
24. Gabriel, R, Eckenberg, R, Paruzynski, A, Bartholomae, CC, Nowrouzi, A, Arens, A *et al.* (2009). Comprehensive genomic access to vector integration in clinical gene therapy. *Nat Med* **15**: 1431–1436.
25. Langmead, B, Trapnell, C, Pop, M and Salzberg, SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.21–R25.10.
26. Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
27. Quinlan, AR and Hall, IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
28. Hematti, P, Hong, BK, Ferguson, C, Adler, R, Hanawa, H, Sellers, S *et al.* (2004). Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol* **2**: e423.
29. Cohen, EAK. Multi-wavelet coherence for point processes on the real line. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Florence, Italy. pp. 2649–2653.
30. Mitra, P and Bokil, H. *Observed Brain Dynamics*. Oxford Universty Press, New York, 2008.
31. Karolchik, D, Hinrichs, AS, Furey, TS, Roskin, KM, Sugnet, CW, Haussler, D *et al.* (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**(Database issue): D493–D496.
32. Brady, T, Agosto, LM, Malani, N, Berry, CC, O'Doherty, U and Bushman, F (2009). HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS* **23**: 1461–1471.
33. Maruggi, G, Porcellini, S, Facchini, G, Perna, SK, Cattoglio, C, Sartori, D *et al.* (2009). Transcriptional enhancers induce insertional gene deregulation independently from the vector type and design. *Mol Ther* **17**: 851–856.
34. Fisher, RA (1922). On the interpretation of chi-squared from contingency tables, and the calculation of *P*. *J Roy Stat Soc* **85**: 87–94.
35. Kent, WJ, Sugnet, CW, Furey, TS, Roskin, KM, Pringle, TH, Zahler, AM *et al.* (2002). The human genome browser at UCSC. *Genome Res* **12**: 996–1006.
36. Tukey, JW. *Exploratory Data Analysis*. Addison-Wesley: Reading, Massachusetts, 1977.

Supplementary Information accompanies this paper on the *Molecular Therapy—Methods & Clinical Development* website (http://www.nature.com/mtm)