Published online 16 March 2011

Nucleic Acids Research, 2011, Vol. 39, No. 11 ***e72***
doi:10.1093/nar/gkr140

A method to sequence and quantify DNA integration for monitoring outcome in gene therapy

Troy Brady[1], Shoshannah L. Roth[1], Nirav Malani[1], Gary P. Wang[1], Charles C. Berry[2], Philippe Leboulch[3,4,5], Salima Hacein-Bey-Abina[6], Marina Cavazzana-Calvo[6], Eirini P. Papapetrou[7,8], Michel Sadelain[7,8], Harri Savilahti[9] and Frederic D. Bushman[1,*]

[1]Department of Microbiology, University of Pennsylvania School of Medicine, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA, [2]Department of Family/Preventive Medicine, University of California, San Diego School of Medicine, San Diego, CA 92093-0901, USA, [3]Commissariat à l'énergie atomique et aux énergies alternatives, Institute of Emerging Diseases and Innovative Therapies (iMETI), [4]INSERM U962 and University of Paris XI, 92265 Fontenay-aux-Roses, France, [5]Genetics Division, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, [6]Department of Biotherapy, Hopital Necker-Enfants Malades, Assistance Publique – Hôpitaux de Paris (AP-HP), Université René Descartes, INSERM, Centre d'Investigation Clinique intégré en Biothérapies, Groupe Hospitalier Universitaire Ouest, AP-HP, Paris, France, [7]Center for Cell Engineering, [8]Molecular Pharmacology and Chemistry Program, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA and [9]Division of Genetics and Physiology, Department of Biology, 20014 University of Turku, Turku, Finland

Received October 29, 2010; Revised January 26, 2011; Accepted February 24, 2011

ABSTRACT

Human genetic diseases have been successfully corrected by integration of functional copies of the defective genes into human cells, but in some cases integration of therapeutic vectors has activated proto-oncogenes and contributed to leukemia. For this reason, extensive efforts have focused on analyzing integration site populations from patient samples, but the most commonly used methods for recovering newly integrated DNA suffer from severe recovery biases. Here, we show that a new method based on phage Mu transposition *in vitro* allows convenient and consistent recovery of integration site sequences in a form that can be analyzed directly using DNA barcoding and pyrosequencing. The method also allows simple estimation of the relative abundance of gene-modified cells from human gene therapy subjects, which has previously been lacking but is crucial for detecting expansion of cell clones that may be a prelude to adverse events.

INTRODUCTION

Human gene therapy has been carried out successfully for several diseases (1–8), but adverse events have occurred in which subjects developed leukemia associated with insertion of therapeutic vectors near proto-oncogenes (3,6,9,10). For this reason, it is important to track the location and abundance of different integration sites in cells from gene therapy-treated subjects. Tracking integration sites is also of interest in the use of transposons as insertional mutations in model organisms and in basic studies of integrating genomic parasites (11,12).

In the previously used protocols for integration site recovery in gene therapy, gene-corrected cells were isolated from patients, genomic DNA was purified and samples were typically cleaved with restriction enzymes. The exposed DNA ends were then ligated to adaptor DNAs and samples amplified using PCR with one primer complementary to the adaptor and the other complementary to the vector DNA terminus. Sites of integration were identified by sequencing PCR products from a primer bound to the vector DNA, so that the sequence read extended into the flanking human DNA (13,14). It was noticed, however, that integration sites were not equally recovered by different enzymes, pointing to a recovery bias for this method. Although sites were efficiently recovered when found near restriction cleavage sites, they were difficult to detect when not positioned optimally (15,16). Moreover, implementing restriction enzyme-based methods is complex—restriction enzymes need to be identified, which do not cleave DNA within amplicons of interest, only enzymes lacking CpG dinucleotides in their recognition sites can be used due to biased distributions of CpG in mammalian DNA, and

*To whom correspondence should be addressed. Tel: +1 215 573 8732; Fax: +1 215 573 4856; Email: bushman@mail.med.upenn.edu

large amounts of genomic DNA are needed. Thus, there is intense interest in alternative approaches.

Improved methods are starting to be described, but all are in early stages (15–18). Concerns with available approaches include efficiency of recovery of rare integration sites (16), and the need for large amounts of genomic DNA for analysis (17,18), which is often not available in gene therapy applications.

Here we report a method for recovering sites of integrated DNA using the bacterial transposase MuA to introduce adaptors into genomic DNA to allow PCR amplification. This method is quick and simple, avoids the bias associated with restriction enzymes, recovers integration sites in a near random fashion, and provides a simple measure of cell clonal abundance.

MATERIALS AND METHODS

Cell lines and gene therapy patient samples

293T and SupT1 cells were infected *in vitro* using a VSV-G pseudotyped MLV or HIV vector produced by transfecting 293T cells. To generate HIV-based vectors, cells were transfected with the LTR–GFP cassette plasmid p156RRLsin-PPTCMVGFPWPRE (19), the packaging construct pCMVdeltaR9 (20) and the vesicular stomatitis virus G-producing plasmid pMD.G. VSV-G pseudotyped MLV particles were produced using pMD.G but in combination with the MLV vector segment (pMX-eGFP) and packaging construct pCGP (pCGP, kindly provided by Paul Bates). Human gene therapy samples consisted of PBMCs (β-thalassemia) or CD3+ cells (SCID-X1). Details of gene therapy samples are reported elsewhere (1,2,7,8). The production of induced pluripotent cells containing a defined number of integration sites is reported in (21).

Mu-mediated integration site recovery

A detailed protocol for preparing Mu reactions is available in the Supplementary Report 1 MuSOP. Briefly, reaction buffer, oligonucleotide donor, target DNA and water are mixed on ice followed by addition of purified Mu transposase. Reactions are then incubated at 30°C for 2–4 h after which 2 µl of the reaction is used as input for PCR. Nested PCR was carried out as described (22) except that extension temperatures during cycling were changed to 70°C for the first seven cycles followed by 67°C for the remaining 37 cycles for samples with HIV-derived vectors. For both HIV- and MLV-based reactions, the number of cycles was reduced to 25 for nested PCR. Single round PCR amplification was performed with primers normally used in the nested round of PCR, each primer encoding 454 adaptor sequences and the LTR primer encoding a DNA barcode (15).

Integration site sequencing and analysis

Integration sites were sequenced using 454 FLX platform technology. Sequences reads were trimmed to remove primer sequence and aligned to the human genome (hg18, version 36.1) using BLAT. Sequences were required to have a single best hit with ≥98% identity to the human genome, to align within 3 bp of the beginning of the sequence read and to contain a perfect match to the expected LTR sequence lying downstream of the barcoded LTR primer. Comparisons to genomic features were carried out as described previously (23,24) using a combination of logistic regression and Bayesian model averaging supplemented by the random Forest machine learning algorithm. Gene expression analyses were based on data from 293T cells (25) with expression measured using the Affymetrix HU133 plus 2.0 gene chip array. Measuring sequence conservation at the site of restriction enzyme or Mu cleavage was performed using weblogo [http://weblogo.berkeley.edu/ and (26)]. Statistical details for generating Lorentz curves and recovery probability plots can be found in Supplementary Report 2 Mu Recovery. Sequence data from this study was submitted to Genbank under accession numbers HR819863–HR863973.

RESULTS

We first quantified biases in the detection of integrated DNA using protocols relying on restriction enzyme cleavage of genomic DNA. Analysis of the recoverability of an integration site based on its proximity to the nearest restriction site showed that integrated DNA ~49 bp from a cleavage site was recovered most frequently, and frequency of recovery decreased sharply at longer or shorter distances (Figure 1A). The case of SCID-X1 patient 10 provides an example of complications due to this recovery bias (3,7). In this patient, an integrated vector activated expression of the nearby BMI1 proto-oncogene, which was associated with massive expansion of leukemic cells. When DNA from blood cells of patient 10 was cleaved and analyzed using four different restriction enzymes, only two enzymes allowed efficient recovery of the BMI1 integration site (Figure 1B). In an extreme effort to circumvent these limitations, a study of SCID-X1 gene-corrected patients used up to six different restriction enzymes to analyze each individual sample, but even with the difficulty and expense of this large scale effort, recovery was still significantly biased (7).

The improved method reported here (Figure 1C and D) substitutes MuA-directed transposition *in vitro* for restriction enzyme cleavage and adaptor ligation [for background and reviews of MuA function see (11,27–33)]. An engineered transposon end is used as donor (Figure 1C), which contains (i) binding sites for the MuA transposase; (ii) an adaptor region complementary to PCR primers; and (iii) an amine blocking group at one DNA 3′-end. In the presence of MuA transposase, the oligonucleotide donors become covalently joined to target DNA, allowing convenient installation of primer sites in human genomic DNA. PCR amplification is then carried out using primers complementary to the vector DNA end and the adaptor. Because the adaptor contains a 5′ overhang and an amino-modified 3′-end, amplification must begin within the vector DNA and extend through the adaptor, preventing adaptor-to-adaptor amplification. As little as 100 ng of genomic target DNA can be used.

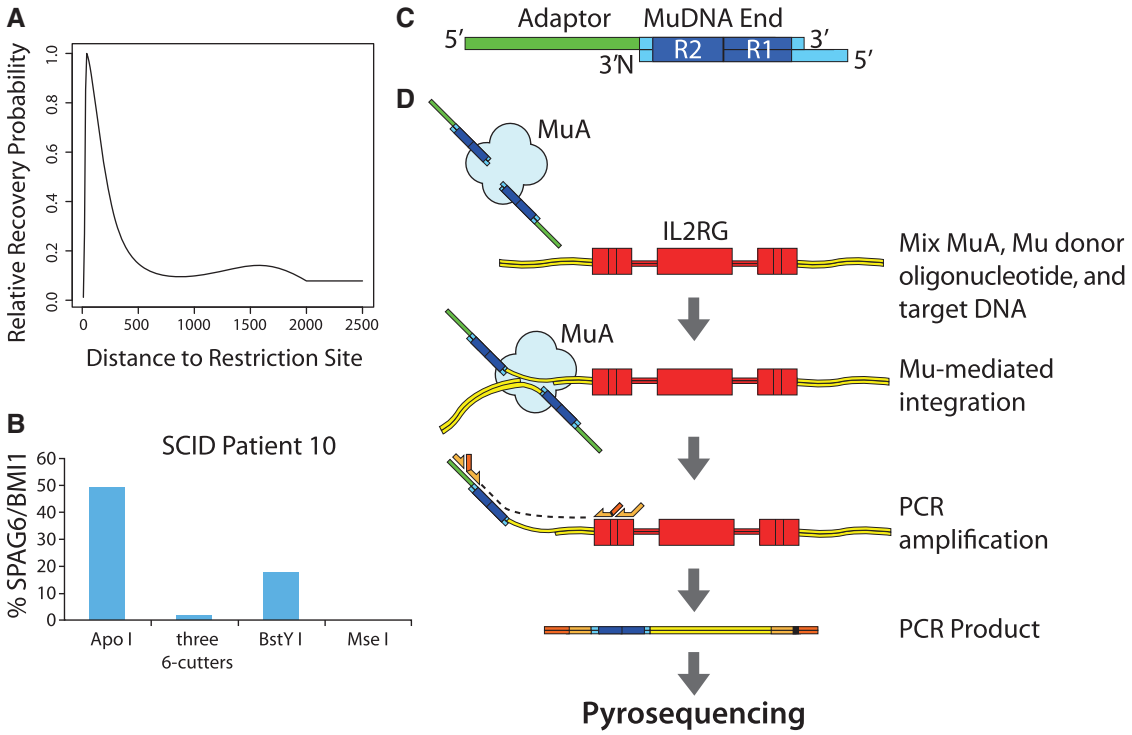

Figure 1. The Mu-based integration site recovery method. (**A**) Severe recovery biases in previous methods using restriction enzyme cleavage. A large collection of integration sites generated from SCID-X1 gene therapy (Supplementary Table S1) were analyzed and plotted to show the relative recovery frequency for different distances between the restriction enzyme site used in genomic DNA cleavage and the vector integration site. The graph summarizes data obtained using six different restriction enzymes. The sharp peak documents the recovery bias (the location of the peak differed modestly for the different restriction enzymes studied; data not shown). (**B**) Biased recovery efficiency for four restriction enzymes in a sample from an adverse event. The integration site within BMI1 was implicated in an adverse event in SCID-X1 patient 10 (3). Each bar indicates the percent of all integration sites from the leukemic cell sample deriving from the BMI1 site for each of the three restriction enzymes or three 6-cutter cocktail (Avr I, Spe I and Nhe I) used for isolation. (**C**) The engineered Mu DNA donor used in these studies. 5′ and 3′ DNA ends are as marked. The 'N' indicates the position of an amino-modifier that blocks the DNA 3′-end to prevent adaptor-to-adaptor amplification. The dark blue indicates binding sites for MuA transposase, light blue a spacer region and green the adaptor sequence for PCR amplification. (**D**) The Mu-mediated integration site recovery method. MuA transposition is used to install the engineered Mu DNA donor (top), allowing PCR amplification (middle). The PCR primers contain DNA barcodes (black segments) and primers for use in 454/Roche pyrosequencing (orange). PCR products can be used directly for pyrosequencing without cloning in bacterial plasmids.

Nested PCR is routinely carried out using the restriction enzyme method, and can be used with the Mu-mediated method. However, tests showed that for samples containing relatively large numbers of integrated vectors, only a single round of PCR was sufficient to yield high quality sequence populations using the Mu-mediated method, and these data sets actually showed improved diversity (described in Supplementary Report 1 'Mu Standard Operating Procedure').

The PCR primers used for the final amplification step are composites, containing both the priming sequences and sequences required for 454/Roche pyrosequencing. Primer sequences also contain a DNA barcode between the 454 sequence and the priming region, allowing large numbers of amplicons to be pooled, sequences determined, then sequence reads parsed using the barcodes (Figure 1D) (34–36). Thus, hundreds of samples can be processed in pools. Pyrosequence reads are then trimmed, aligned to the human genome and distributions analyzed.

To test the Mu-mediated method, we determined 25 194 total integration site sequences, which yielded 3382 unique vector integration sites after condensing duplicates. We analyzed both HIV- and γ-retrovirus-based vectors. Samples studied included cells from patients in two gene therapy trials, which treated SCID-X1 (3) and β-thalassemia (8) and tissue culture cells infected *in vitro* (summarized in Supplementary Table S1).

We first compared recovery biases of the Mu and restriction enzyme-based methods. We determined the sequence preferences for Mu integration in human DNA *in vitro* for 5968 integration site sequence reads that included the Mu-end oligonucleotide DNA. Alignment of human sequences at Mu integration sites revealed a detectable consensus sequence closely resembling that reported previously for Mu transposition (37), but with much lower information content than cleavage sites for restriction enzymes (Figure 2A and B and Supplementary Report 2), indicating less bias in the cleaving/joining reactions.

The performance of the restriction enzyme method was next compared with the Mu-mediated method by measuring recovery biases for each method, then annotating the human genome for calculated recovery rates based on the data. Thus, for any integration site in the human genome,

the likelihood of recovery at each base pair could be calculated for each recovery method (Supplementary Report 2).

Figure 3A illustrates the relative recovery frequencies using the LMO2 promoter as an example, which was chosen because the gene has been involved in several adverse events in SCID-X1 gene therapy (3,38). Note that although Figure 3A summarizes results on a single region of the genome, the biases were measured genome-wide from integration site data for each method and are shown at LMO2 for purposes of illustration. Recovery was relatively consistent for the Mu-mediated method over all sites, whereas the restriction enzyme methods show sharp peaks and valleys, where valleys indicate locations where an integration event would be difficult or impossible to isolate and the peaks frequently recovered sites that would mask more rare sites. Figure 3B shows the data plotted as the cumulative recovery frequency, where perfectly unbiased recovery would be indicated by a curve that followed the diagonal from lower left to upper right. The Mu-mediated recovery method most closely approaches the diagonal and is significantly closer than even the method using six restriction enzymes ($P < 0.001$; see Supplementary Report 2 for statistical methods), documenting that the Mu-mediated method is the least biased.

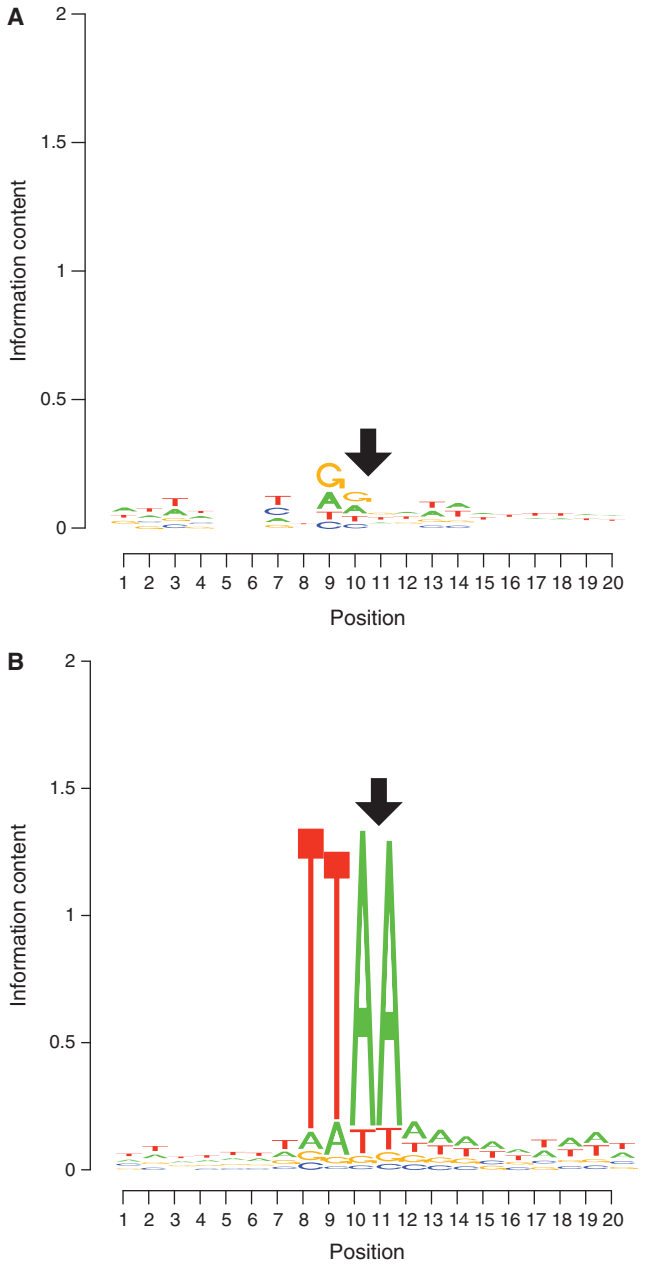

Figure 2. Consensus sequences at points of adaptor addition. (**A**) Information content at junctions between the engineered Mu DNA and human DNA derived from vector integration site sequence reads. The *x*-axis shows the DNA sequence position, where the site of joining to Mu DNA is between positions 10 and 11 (arrow). Perfect sequence conservation has information content of 2 bits (*y*-axis). Note that some bases have little or no information content, so no letters are visible. (**B**) Information content at adaptor junctions from vector integration sites recovered after cleavage with the restriction enzyme Mse I, where the site of cleavage is between positions 10 and 11 (arrow).

Another means of quantifying recovery biases involves comparing the increased sampling effort required to reach the results of a perfectly unbiased method. The slight biases introduced by the Mu method would require only a 10% increase in sampling effort to achieve the efficiency of a perfectly unbiased method (Supplementary Report 2). In contrast, for some restriction enzyme methods, the needed increase is too large to measure accurately (>50-fold). For the pool of all six restriction enzymes, a 45% increase in effort would be needed. Thus the Mu-mediated method yields less biased recovery with much less effort than any form of restriction enzyme-based method.

We next investigated the experimental effort required to recover all members of a fully defined integration site population using the Mu-based method. We prepared an induced pluripotent stem (iPS) cell line in which all cells contained the same six known lentiviral vector integration sites. We carried out 12 independent integration site recovery reactions using the Mu-mediated method, recovering an average of 1068 sequence reads per replicate. We found that six of the 12 replicate reactions recovered all six sites. The replicates with fewer than six yielded either four or five of the sites. To assess the sampling effort required for recovery of all six sites, we pooled different numbers of Mu reactions computationally and assessed recovery. For pools of three of the 12 Mu reactions, 98% contained all six sites. For pools of five Mu reactions, 100% of the pools contained all six sites. Thus, 3–5 independent Mu reactions are enough to completely sample an integration site population of this size at the sequencing depth used. For comparison, three restriction enzymes were tested for recovery of the six sites, and only one yielded all six.

The genome-wide distribution patterns of integration target sites for HIV and γ-retroviruses have been studied extensively (13,14,23,39,40), allowing us to assess whether the Mu-mediated method reported similar trends. For studies of integration frequency near genomic landmarks, restriction enzyme-based methods usually provided an adequate overview, because the restriction biases are only weakly related to those landmarks. Figure 4A and B compares the distributions of integration sites isolated using the two methods for HIV and γ-retrovirus-based

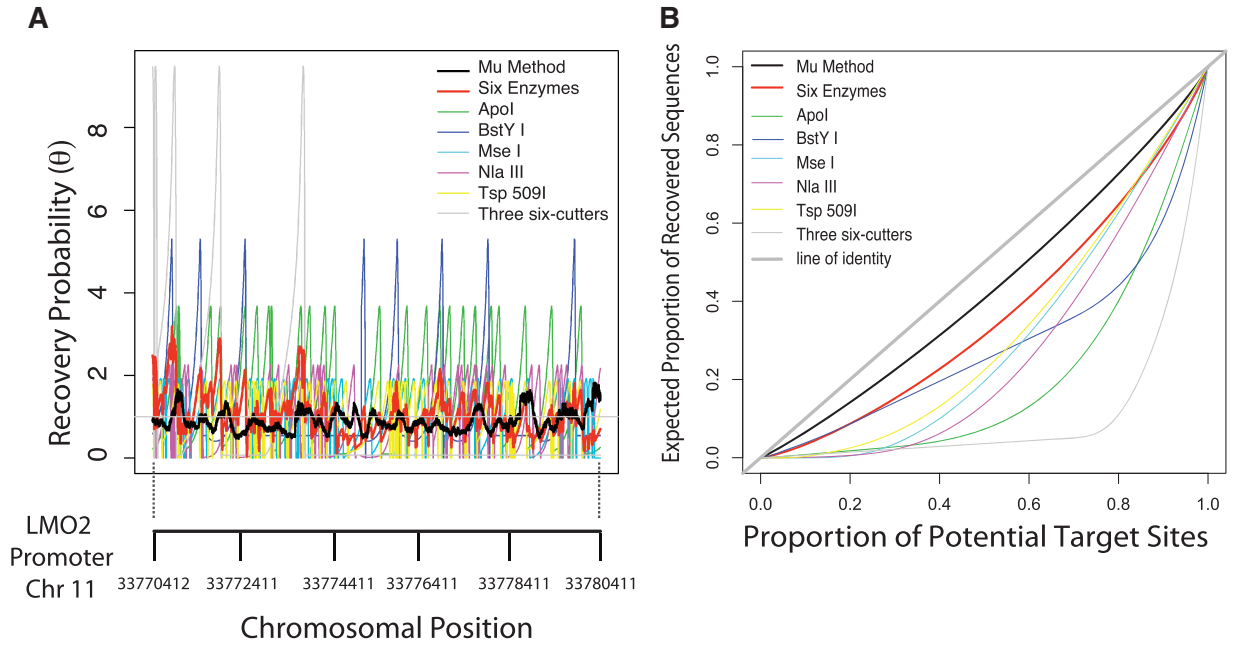

Figure 3. Reduced recovery bias using the Mu-mediated integration site isolation method. (**A**) Relative recovery rates compared at the LMO2 promoter for (i) the Mu-mediated method; (ii) six tests with single restrictions enzymes or pools or (iii) a mixture of all six. The recovery rates were calculated from data on the placement of integration sites relative to restriction enzyme cleavage sites or Mu transposase sites used in their isolation, using the data in Supplementary Table S1 and statistical methods described in Supplementary Report 2. Calculated integration site recovery rates were then used to annotate each base over 10 kb at the LMO2 promoter (chr11, bases 33770412–33780411). 'Six enzymes' indicates pooled data for the six sets below. For a perfectly unbiased method, all such rates equal 1.0. (**B**) Statistical analysis of biases in recovery of integration sites. The *x*-axis plots each base of the LMO2 promoter analyzed above, treating each as a potential integration target. Sites were ranked by expected ease of isolation, with the easiest to isolate to the right. The *y*-axis shows the calculated proportion of sequences recovered given the measured recovery biases. Perfect unbiased recovery would follow a line from lower left to upper right. Statistical methods and *P* values for pair-wise comparisons are summarized in Supplementary Report 2.

vectors. Integration sites are compared from infections of tissue culture cells and for one SCID-X1 patient (3,7). Figure 4A summarizes the relationship of integration sites to genomic features, using a heat map format to indicate increased or decreased integration frequency compared with random distributions. For both HIV and γ-retroviral vectors, the genome-wide trends were closely similar for the Mu-mediated or restriction enzyme-mediated methods. Both HIV-based and γ-retrovirus-based vectors favor integration in regions of high gene density and associated genomic landmarks. Gamma-retroviruses favor integration near gene 5′-ends. HIV and γ-retroviruses show a complex pattern of favored and disfavored integration sites near regions of histone methylation, acetylation and bound proteins, as indicated by comparison to data from ChIP-seq experiments (41–44) (Figure 4B), again showing favored integration near marks of active transcription. These patterns matched closely for the Mu-mediated and restriction-enzyme-based methods. The genome-wide patterns for SCID-X1 gene therapy closely paralleled those seen for γ-retroviral vectors, as reported previously (3,6,7). Thus, conclusions on genome-wide distributions of integration sites from the Mu-mediated method parallel those from extensive previous studies.

The Mu-mediated method also allows a new approach to quantifying the relative frequency of gene-corrected cells in patient samples. Gene corrected cell clones that are present in many copies will contribute a relatively larger proportion of their integration site DNA to the genomic DNA pool after purification, providing an increased number of target sites for Mu integration *in vitro*. As a result, relatively larger numbers of independent Mu integration events will lead to recovery of the same high abundance vector integration site. Quantifying the number of independent Mu integration sites per vector integration site therefore provides a simple measure of the relative abundance.

We tested this in a sample from a β-thalassemia gene therapy trial, in which a cell harbouring a single vector integration site in the HMGA2 gene expanded to comprise more than one-third of the gene-corrected peripheral blood mononuclear cells (PBMC) population, as documented by quantitative PCR assays (8). We recovered 10 independent Mu integration sites for the HMGA2 vector integration site, while all other vector integration sites were recovered with only a single Mu site (Figure 4C). This finding makes the important clinical point that the HMGA2 site is the only high abundance integration site in the expanded cell clone. Using the restriction enzyme cleavage method for integration site recovery, only two out of three of the enzymes used allowed isolation of the HMGA2 site (data not shown).

Lastly, we present a method for suppressing PCR contamination. In assessing collections of gene therapy samples, multiple tubes are commonly processed in parallel using nested PCR, providing an ideal setting for migration of PCR products between samples. In practice controlling

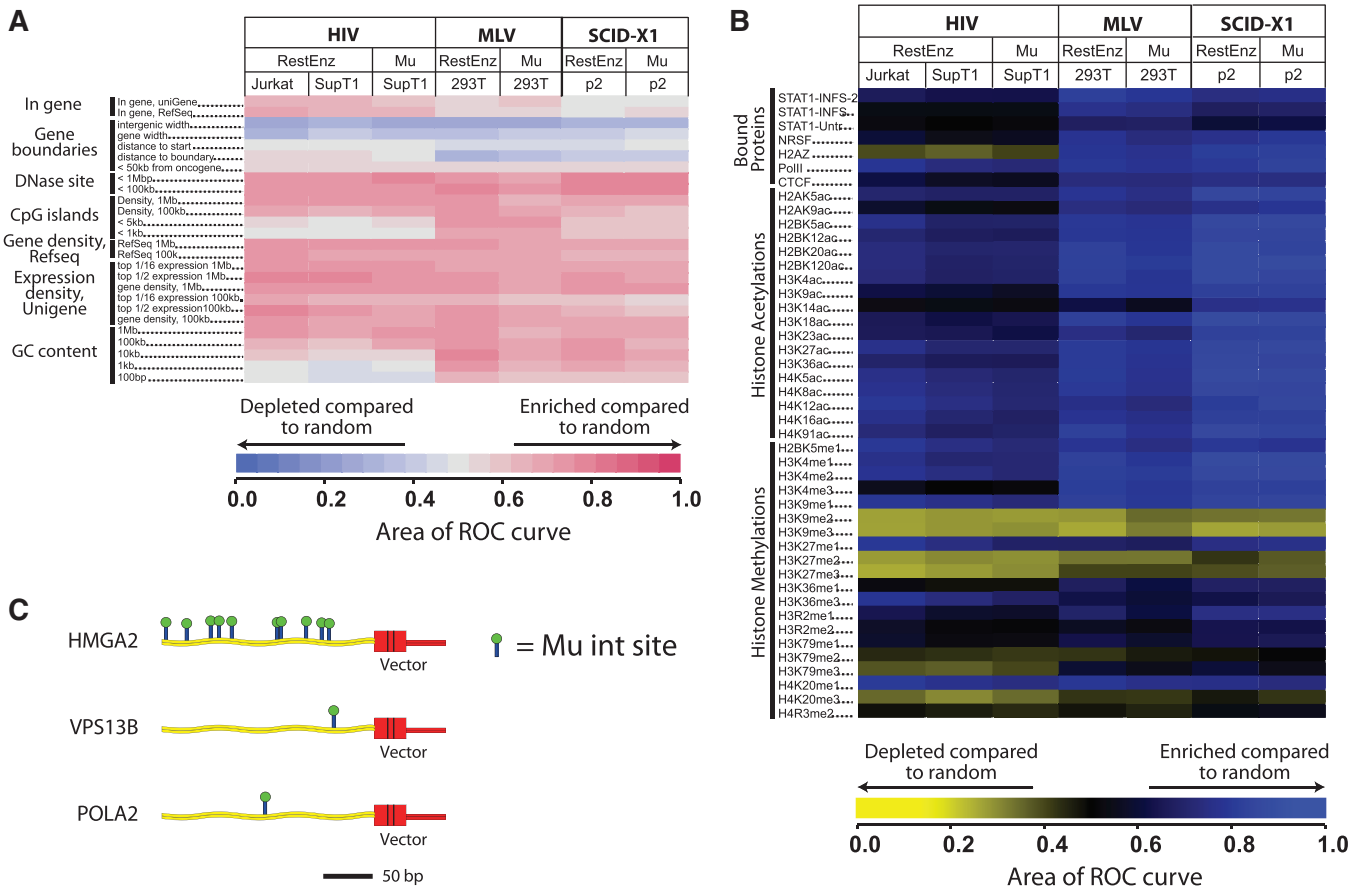

Figure 4. Mu-mediated integration site recovery. (**A**) Integration frequency near genomic land marks. (**B**) Integration frequency near sites of histone methylation, acetylation or bound chromosomal proteins. Data sets compared are indicated by the column headings, features analyzed by the rows. Heat maps were constructed using the receiver operating characteristic (ROC) area method to compare the observed distributions to random distributions of integration sites (23,45). ROC areas of 0.5 indicated integration sites are present near the indicated genomic features as often as expected by chance. ROC areas >0.5 indicate positive association and areas <0.5 negative association. Associations are colour coded as indicated by the key at the bottom of each map. In (A), for the Gene Density, Expression Density and GC content measures, several different length genomic intervals were used for comparisons, which are indicated by numbers to the right of the black bar. In (B), ChIP-seq analysis was used to map the genome-wide distributions of sites of histone post-translational methylation or acetylation, or bound DNA binding proteins (41–44), and the results compared with integration site distributions. Statistical analysis shows that most associations where discernable colour can be seen in a heat-map tile achieve statistical significance. (**C**) Comparisons of the numbers and positions of Mu integration sites that allowed recovery of the vector integration sites at HMGA2, VPS13B and POLA2 generated during gene therapy for β-thalassemia.

contamination is challenging even for experienced experimentalists. We have devised a method that suppresses this, in which separate adaptors are used for each sample, so that contaminating PCR products are not amplifiable outside the correct PCR reaction. In reconstruction experiments, this has been effective at suppressing cross-over in our laboratory (data not shown). Use of multiple adaptors is described in the Standard Operating Procedure in the Supplementary Data for this article.

DISCUSSION

In summary, the Mu-mediated integration site recovery method allows simplified recovery of integration sites and estimation of relative abundance. Use of protocols based on single restriction enzymes results in failure to isolate integration sites that are not near restriction enzyme recognition sites (7,15,16), which can be crucial in monitoring adverse events during gene therapy. The challenges posed by biased isolation have been underestimated in some of the early literature in this field. One previous study attempted to circumvent recovery biases by using six restriction enzymes to study single samples, but even with this added complication and expense, recovery using this method is still more biased than with the Mu-based method (Figure 3B; $P < 0.001$). Thus in cases where it is critical to use integration site data to identify expanded cell clones, the Mu method is attractive and convenient.

Methods for deep sequencing are in a state of rapid transition. The Mu-mediated method described here can in principal be adapted to any of the next generation sequencing platforms. At this writing, the Solexa/Illumina method is least expensive per base, but potentially inconvenient because analysis of a handful of samples will often take up only a fraction of a run, requiring complicated

coordination with others to fill out a run. The 454/Roche method is more expensive per base, but the availability of a benchtop instrument for smaller runs (the 'Junior' instrument http://www.gsjunior.com/) simplifies throughput. It is highly likely that additional sequencing methods will become available in the near future that may also be useable with the Mu-mediated method.

The demands placed on integration site recovery technology vary by disease state. The frequency of gene corrected cells varies from 100% (T cells in SCID-X1) to <1% [early adenosine deaminase (ADA) studies], so the demands on the technology differ for different diseases. For very low level clones, recovering integration sites and estimating their abundance is at present challenging for any technology.

Additional methods are starting to be proposed for integration site analysis, including methods based on DNA shearing (N. Gillet, N. Malani, N. Gormley, R. Carter, A. Melamed, D. Bentley, C. Berry, F. Bushman, G. Taylor and C. Bangham, submitted for publication) or limited extension from integrated vectors with a DNA polymerase followed by RNA ligation ('nrPCR') (16). Each of these methods is of interest but each may have inefficient steps. For the Mu-mediated method, it can be challenging to obtain enough Mu integration events to query the full human genome efficiently, though the method is suitable for analysis of large numbers of samples with small amounts of starting genomic DNA. Ongoing use has shown the method to be effective in practice. For DNA shearing, it can be challenging to obtain efficient ligation after repairing broken DNA ends, and to work with small amounts of DNA. For the nrPCR method, efficiency may be an issue (16).

Two reports in the peer-reviewed literature document the utility of the Mu-mediated method. In one case, oligoclonal reconstitution during lentiviral vector-mediated gene correction was documented using Mu-mediated recovery of integration sites during β-thalassemia gene correction in mice (46). In this case, the inferred rank order of integration site abundance from sequence read counts was similar to that inferred by quantifying the number of independent Mu-transposition events *in vitro* that led to site recovery. The congruence of these two measures supports the idea that quantification relative to clonal abundance was consistent. In the second study (21) the Mu-mediated method was used along with other methods and found to be comparably efficient. Long term, it will be useful to compare the effectiveness and convenience of present and future methods for quantifying integration site abundance on defined integration site populations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to members of the Bushman laboratory for help and suggestions.

FUNDING

Funding for open access charge: National Institute of Health (grants AI52845 and AI082020); New York State Stem Cell Science NYSTEM (N08T-060) University of Pennsylvania Center for AIDS Research and the Penn Genome Frontiers Institute with a grant with the Pennsylvania Department of Health. T.B. is a Special Fellow of the Leukemia & Lymphoma Society.

Conflict of interest statement. The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

REFERENCES

1. Cavazzana-Calvo,M., Hacein-Bey,S., de Saint Basile,G., Gross,F., Yvon,E., Nusbaum,P., Selz,F., Hue,C., Certain,S., Casanova,J.L. *et al.* (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, 669–672.
2. Hacein-Bey-Abina,S., Le Deist,F., Carlier,F., Bouneaud,C., Hue,C., De Villartay,J.P., Thrasher,A.J., Wulffraat,N., Sorensen,R., Dupuis-Girod,S. *et al.* (2002) Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N. Engl. J. Med.*, **346**, 1185–1193.
3. Hacein-Bey-Abina,S., Garrigue,A., Wang,G.P., Soulier,J., Lim,A., Morillon,E., Clappier,E., Caccavelli,L., Delabesse,E., Beldjord,K. *et al.* (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.*, **118**, 3132–3142.
4. Aiuti,A., Slavin,S., Aker,M., Ficara,F., Deola,S., Mortellaro,A., Morecki,S., Andolfi,G., Tabucchi,A., Carlucci,F. *et al.* (2002) Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science*, **296**, 2410–2413.
5. Ott,M.G., Schmidt,M., Schwarzwaelder,K., Stein,S., Siler,U., Koehl,U., Glimm,H., Kuhlcke,K., Schilz,A., Kunkel,H. *et al.* (2006) Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat. Med.*, **12**, 401–409.
6. Deichmann,A., Hacein-Bey-Abina,S., Schmidt,M., Garrigue,A., Brugman,M.H., Hu,J., Glimm,H., Gyapay,G., Prum,B., Fraser,C.C. *et al.* (2007) Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J. Clin. Invest.*, **117**, 2225–2232.
7. Wang,G.P., Berry,C.C., Malani,N., Leboulch,P., Fischer,A., Hacein-Bey-Abina,S., Cavazzana-Calvo,M. and Bushman,F.D. (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.
8. Cavazzana-Calvo,M., Payen,E., Negre,O., Wang,G., Hehir,K., Fusil,F., Down,J., Denaro,M., Brady,T., Westerman,K. *et al.* (2010) Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature*, **467**, 318–322.
9. Hacein-Bey-Abina,S., von Kalle,C., Schmidt,M., Le Deist,F., Wulffraat,N., McIntyre,E., Radford,I., Villeval,J.L., Fraser,C.C., Cavazzana-Calvo,M. *et al.* (2003) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.*, **348**, 255–256.
10. Hacein-Bey-Abina,S., Von Kalle,C., Schmidt,M., McCormack,M.P., Wulffraat,N., Leboulch,P., Lim,A., Osborne,C.S., Pawliuk,R., Morillon,E. *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.
11. Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (2002) *Mobile DNA II*. ASM Press, Washington, DC, USA.
12. Bushman,F.D. (2001) *Lateral DNA Transfer: Mechanisms and Consequences*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.
13. Schroder,A.R., Shinn,P., Chen,H., Berry,C., Ecker,J.R. and Bushman,F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521–529.

14. Wu,X., Li,Y., Crise,B. and Burgess,S.M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.
15. Wang,G.P., Garrigue,A., Ciuffi,A., Ronen,K., Leipzig,J., Berry,C., Lagresle-Peyrou,C., Benjelloun,F., Hacein-Bey-Abina,S., Fischer,A. *et al*. (2008) DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res*., **36**, e49.
16. Gabriel,R., Eckenberg,R., Paruzynski,A., Bartholomae,C.C., Nowrouzi,A., Arens,A., Howe,S.J., Recchia,A., Cattoglio,C., Wang,W. *et al*. (2009) Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med*., **15**, 1431–1436.
17. Langridge,G.C., Phan,M.D., Turner,D.J., Perkins,T.T., Parts,L., Haase,J., Charles,I., Maskell,D.J., Peters,S.E., Dougan,G. *et al*. (2009) Simultaneous assay of every salmonella typhi gene using one million transposon mutants. *Genome Res*., **19**, 2308–2316.
18. Williams-Carrier,R., Stiffler,N., Belcher,S., Kroeger,T., Stern,D.B., Monde,R.A., Coalter,R. and Barkan,A. (2010) Use of illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy mutator lines of maize. *Plant J*., **63**, 167–177.
19. Follenzi,A., Ailes,L.E., Bakovic,S., Gueuna,M. and Naldini,L. (2000) Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. *Nat. Genet*., **25**, 217–222.
20. Naldini,L., Blomer,U., Gallay,P., Ory,D., Mulligan,R., Gage,F.H., Verma,I.M. and Trono,D. (1996) In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science*, **272**, 263–267.
21. Papapetrou,E.P., Lee,G., Malani,N., Setty,M., Riviere,I., Tirunagari,L.M., Kadota,K., Roth,S.L., Giardina,P., Viale,A. *et al*. (2011) Genomic safe harbors permit high beta-globin transgene expression in thalassemia induced pluripotent stem cells. *Nat. Biotechnol*., **29**, 73–78.
22. Ciuffi,A., Ronen,K., Brady,T., Malani,N., Wang,G., Berry,C.C. and Bushman,F.D. (2009) Methods for integration site distribution analyses in animal cell genomes. *Methods*, **47**, 261–268.
23. Berry,C., Hannenhalli,S., Leipzig,J. and Bushman,F.D. (2006) Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol*., **2**, e157.
24. Brady,T., Lee,Y.N., Ronen,K., Malani,N., Berry,C.C., Bieniasz,P.D. and Bushman,F.D. (2009) Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev*., **23**, 633–642.
25. Ciuffi,A., Llano,M., Poeschla,E., Hoffmann,C., Leipzig,J., Shinn,P., Ecker,J.R. and Bushman,F. (2005) A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med*., **11**, 1287–1289.
26. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res*., **14**, 1188–1190.
27. Mizuuchi,K. (1992) Polynucleotidyl transfer reactions in transpositional DNA recombination. *J. Biol. Chem*., **287**, 21273–21276.
28. Chaconas,G. (1999) Studies on a "jumping gene machine": Higher-order nucleoprotein complexes in mu DNA transposition. *Biochem. Cell Biol*., **77**, 487–491.
29. Paatero,A.O., Turakainen,H., Happonen,L.J., Olsson,C., Palomaki,T., Pajunen,M.I., Meng,X., Otonkoski,T., Tuuri,T., Berry,C. *et al*. (2008) Bacteriophage mu integration in yeast and mammalian genomes. *Nucleic Acids Res*., **36**, e148.
30. Savilahti,H., Rice,P.A. and Mizuuchi,K. (1995) The phage mu transpososome core: DNA requirements for assembly and function. *EMBO J*., **14**, 4893–4903.
31. Haapa,S., Taira,S., Heikkinen,E. and Savilahti,H. (1999) An efficient and accurate integration of mini-mu transposons *in vitro*: A general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res*., **27**, 2777–2784.
32. Pajunen,M., Poussu,E., Turakainen,H. and Savilahti,H. (2009) Application of mu *in vitro* transposition for high-precision mapping of protein-protein interfaces on a yeast two-hybrid platform. *Methods*, **49**, 255–262.
33. Wei,S.Q., Mizuuchi,K. and Craigie,R. (1997) A large nucleoprotein assembly at the ends of the viral DNA mediates retroviral DNA integration. *EMBO J*., **16**, 7511–7520.
34. Hoffmann,C., Minkah,N., Leipzig,J., Wang,G., Arens,M.Q., Tebas,P. and Bushman,F.D. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res*., **35**, e91.
35. Binladen,J., Gilbert,M.T., Bollback,J.P., Panitz,F., Bendixen,C., Nielsen,R. and Willerslev,E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.
36. Hamady,M., Walker,J.J., Harris,J.K., Gold,N.J. and Knight,R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, **5**, 235–237.
37. Haapa-Paananen,S., Rita,H. and Savilahti,H. (2002) DNA transposition of bacteriophage mu. A quantitative analysis of target site selection *in vitro*. *J. Biol. Chem*., **277**, 2843–2851.
38. Howe,S.J., Mansour,M.R., Schwarzwaelder,K., Bartholomae,C., Hubank,M., Kempski,H., Brugman,M.H., Pike-Overzet,K., Chatters,S.J., de Ridder,D. *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.
39. Aiuti,A., Cassani,B., Andolfi,G., Mirolo,M., Biasco,L., Recchia,A., Urbinati,F., Valacca,C., Scaramuzza,S., Aker,M. *et al*. (2007) Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J. Clin. Invest*., **117**, 2233–2240.
40. Hematti,P., Hong,B.K., Ferguson,C., Adler,R., Hanawa,H., Sellers,S., Holt,I.E., Eckfeldt,C.E., Sharma,Y., Schmidt,M. *et al*. (2004) Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol*., **2**, e423.
41. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
42. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al*. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet*., **40**, 897–903.
43. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
44. Wang,Z., Zang,C., Cui,K., Schones,D.E., Barski,A., Peng,W. and Zhao,K. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.
45. Brady,T., Agosto,L.M., Malani,N., Berry,C.C., 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.
46. Ronen,K., Negre,O., Roth,S., Malani,N., Brady,T., Denaro,M., Fusil,F., Gillet-Legrand,B., Hehir,K., Beuzard,Y. *et al*. (2011) Distribution of lentiviral vector integration sites in mice following therapeutic gene transfer to treat beta-thalassemia. *Blood*, in press.