Title

INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions

Authors

Berry, Charles C
Nobles, Christopher
Six, Emmanuelle
et al.

Publication Date

2017-03-01

DOI

10.1016/j.omtm.2016.11.003

Peer reviewed

Molecular Therapy
Methods & Clinical Development
Original Article

INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions

Charles C. Berry,[1,7] Christopher Nobles,[2,7] Emmanuelle Six,[3,4,7] Yinghua Wu,[2,7] Nirav Malani,[2,7] Eric Sherman,[2,7] Anatoly Dryga,[2,7] John K. Everett,[2] Frances Male,[2] Aubrey Bailey,[2] Kyle Bittinger,[2] Mary J. Drake,[2] Laure Caccavelli,[5,6] Paul Bates,[2] Salima Hacein-Bey-Abina,[5,6] Marina Cavazzana,[5,6] and Frederic D. Bushman[2]

[1]Department of Family Medicine and Public Health, UC San Diego, La Jolla, CA 92093, USA; [2]Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-6076, USA; [3]Paris Descartes-Sorbonne Paris Cité University, Imagine Institute, 75015 Paris, France; [4]INSERM 24, Laboratory of Human Lymphohematopoiesis, 75015 Paris, France; [5]Biotherapy Department, Necker Children's Hospital, Assistance Publique-Hôpitaux de Paris, 75014 Paris, France; [6]Biotherapy Clinical Investigation Center, Groupe Hospitalier Universitaire Ouest, Assistance Publique-Hôpitaux de Paris, INSERM, 75014 Paris, France

Analysis of sites of newly integrated DNA in cellular genomes is important to several fields, but methods for analyzing and visualizing these datasets are still under development. Here, we describe tools for data analysis and visualization that take as input integration site data from our INSPIIRED pipeline. Paired-end sequencing allows inference of the numbers of transduced cells as well as the distributions of integration sites in target genomes. We present interactive heatmaps that allow comparison of distributions of integration sites to genomic features and that support numerous user-defined statistical tests. To summarize integration site data from human gene therapy samples, we developed a reproducible report format that catalogs sample population structure, longitudinal dynamics, and integration frequency near cancer-associated genes. We also introduce a novel summary statistic, the UC50 (unique cell progenitors contributing the most expanded 50% of progeny cell clones), which provides a single number summarizing possible clonal expansion. Using these tools, we characterize ongoing longitudinal characterization of a patient from the first trial to treat severe combined immunodeficiency-X1 (SCID-X1), showing successful reconstitution for 15 years accompanied by persistence of a cell clone with an integration site near the cancer-associated gene CCND2. Software is available at https://github.com/BushmanLab/INSPIIRED.

INTRODUCTION

Retroviruses, transposons, and other mobile DNA elements directly integrate their DNA into the chromosomes of host cells.[1–8] Distributions of newly integrated DNA elements can be characterized using next-generation sequencing, as is described in the companion paper in this issue of *MolecularTherapy: Methods & Clinical Development*[9] and in many previous studies (e.g., Bushman,[1] Craig et al.,[2] Schröder et al.,[3] Mitchell et al.,[4] Maldarelli et al.,[6] Cohn et al.,[8] Wu et al.,[10] Hoffmann et al.,[11] and Biffi et al.[12]). Contemporary methods take advantage of Illumina paired-end sequencing to report the location of newly integrated DNA.[13–16] Multiple reports have described methods for quantifying and analyzing such data, but optimal methods for statistical analysis, data reduction, and data visualization are the topics of ongoing development.[13,14,16–19] Here, we describe a suite of tools for integration site analysis and visualization that are useful for characterizing samples from human gene therapy and other applications.

In the case of gene modification in circulating blood cells, it is possible to sample cell populations from blood longitudinally and sequence sites of integration of the gene-correcting vector.[20–27] An important question centers on how best to quantify the numbers and types of gene-modified cell clones contributing blood cells to the periphery. For example, adverse events have been reported where expanded cell clones in blood became frank leukemia,[28–31] and this can be tracked using quantitative integration site data. Complicating the analysis, simply counting the number of integration site sequence reads does not accurately report clonal abundance because of distortions resulting from PCR steps in the integration site recovery procedure.[32,33]

We have previously described a method for abundance estimation based on paired-end sequencing of PCR products containing integration site sequences that allows for accurate quantification of gene-modified cells.[34] DNA is sheared using sonication, DNA linkers are ligated to free DNA ends, and then samples are amplified using primers complementary to the integrated vectors and ligated linkers. Genomic sequence information is acquired from both the linker end and the integrated vector end. For the case of an expanded clone, many different DNA breaks and sites of linker ligation are associated with the unique integration site from the expanded clone. This allows for the estimation of abundance using the number of different linker ligation sites as a surrogate for the number of cells sampled. We have

Received 20 August 2016; accepted 15 November 2016;
http://dx.doi.org/10.1016/j.omtm.2016.11.003.

[7]These authors contributed equally to this work.

Correspondence: Frederic D. Bushman, Department of Microbiology, University of Pennsylvania Perelman School of Medicine, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA.
E-mail: bushman@mail.med.upenn.edu

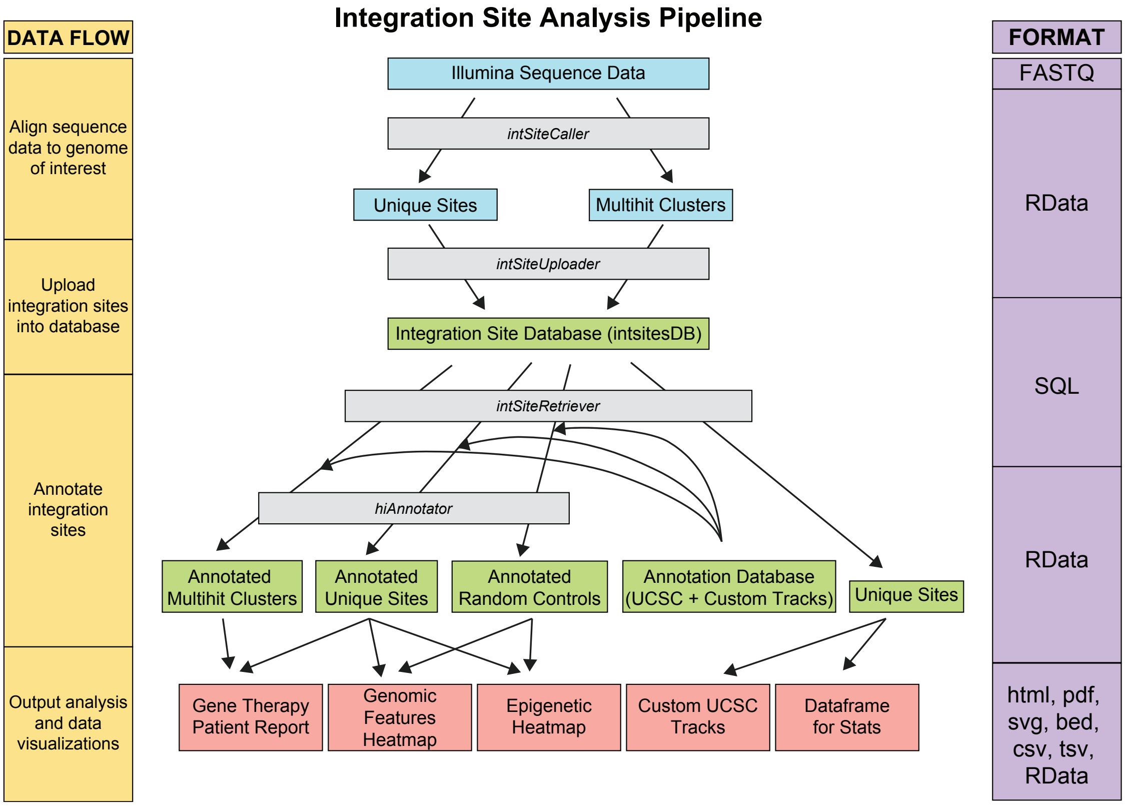
Integration Site Analysis Pipeline
DATA FLOW
Align sequence data to genome of interest
Upload integration sites into database
Annotate integration sites
Output analysis and data visualizations
Illumina Sequence Data
intSiteCaller
Unique Sites
Multihit Clusters
intSiteUploader
Integration Site Database (intsitesDB)
intSiteRetriever
hiAnnotator
Annotated Multihit Clusters
Annotated Unique Sites
Annotated Random Controls
Annotation Database (UCSC + Custom Tracks)
Unique Sites
Gene Therapy Patient Report
Genomic Features Heatmap
Epigenetic Heatmap
Custom UCSC Tracks
Dataframe for Stats
FORMAT
FASTQ
RData
SQL
RData
html, pdf, svg, bed, csv, tsv, RData

Figure 1. Diagram of the INSPIIRED Pipeline

File types generated at each step are indicated on the right. RData files of Unique Sites, Multihit Clusters, and Annotated objects contain GRanges objects from the Bioconductor GenomicRanges R package.

published statistical tools for analysis of such data and applied these tools to track several gene therapy trials.[20,21,25,26,34–39] Other groups have also used related methods.[6,8,24,40–44]

Here, we present tools for integration site sequence analysis and the quantification of clonal abundance, and describe applications in human gene therapy. These methods can also be used for tracking latently infected cells in HIV-positive subjects and monitoring experiments using insertional mutagens, and in mechanistic studies of DNA integration. We describe a heatmap format for the analysis of relationships among integration site distributions, genomic features, and sites of epigenetic modification. These analyses allow users to carry out numerous custom statistical comparisons with annotations, random distributions, and other datasets by simply pointing and clicking. We also present a series of analytical tools for use with patient samples to characterize integration site population structure and possible adverse events. Results are packaged into reproducible reports (html or pdf file format), allowing for version tracking of the code, datasets used, and external datasets queried. Using these tools, we describe examples of tracking a subject from the first gene therapy trial to treat severe combined immunodeficiency-X1 (SCID-X1) deficiency. The data demonstrate durable reconstitution accompanied by a clonal expansion of cells harboring an integrated vector near the cancer-associated gene CCND2.[25]

RESULTS

The INSPIIRED Pipeline

The INSPIIRED pipeline is summarized in Figure 1. The first steps involve the generation of a sequence library and sequence data acquisition, genomic alignment, analysis of viral integration sites in repeated sequences, and quality controls (described in the accompanying paper[9]). The intSiteCaller program takes FASTQ files as input. After sequence quality filtering, trimming of DNA sequences added during library construction, and sequence alignment, unique sites and integration sites in repeated sequences (“multihits”) are saved in an intermediary binary format (RData file format). All sites from each run of the sequencing instrument are uploaded into a MySQL database (IntSiteDB) for storage and for use in downstream analysis using a utility script (intSiteUploader). Alternatively, the INSPIIRED pipeline also supports use of a SQLite database, which is provided with the software. The IntSiteDB stores genomic locations of integration sites together with PCR break points and their counts, which is used for estimation of abundance.[34] All downstream analyses are carried out using genomic locations and sonic break points. The IntSiteDB schema is shown in Figure S1.

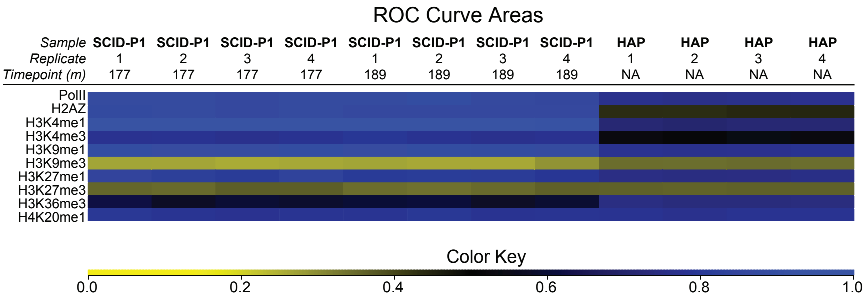

Figure 2. Heatmap Summarizing the Density of Integration Sites versus the Density of Sites of Epigenetic Modification on the Human Genome

Integration site distributions are compared with the density within a 10 kb window of epigenetic marks mapped in CD133$^+$ progenitor cells. Samples are shown in the columns, and bound proteins recovered by ChIP-seq are shown in rows. Associations are quantified using the ROC area method. The values of ROC areas are shown in the color key at the bottom. ChIP-seq data are from Raney et al.[49]

We use intSiteRetreiver, a key component of the INSPIIRED software package, to retrieve unique sites and multihits for chosen samples from the IntSiteDB for analysis. The database organization allows the combined analysis of multiple samples from different instrument runs. Integration sites are annotated with the hiAnnotator R software package (http://bioconductor.org/packages/release/bioc/html/hiAnnotator.html), which makes use of genomic features compiled by the UCSC Genome Bioinformatics Group.

For studies of gene therapy subjects, patient metadata and specimen information are stored in an accompanying gene therapy specimen management database that includes anonymized patient identifiers, cell types analyzed, and time point data. These features are then used in the gene therapy patient reports described below. The separation of the pipeline into several output products (patient report and heatmaps) and databases (integration site, patient metadata, and annotation) provides flexibility in development and use.

Analysis of the Relationship between Integration Site Locations and Genomic Annotation

We use heatmaps to summarize the relationships between integration site distributions and genomic annotations (Figures 2, 3, and 4). These maps were introduced in Berry et al.,[35] which presents more background and examples of their uses. The heatmaps summarize information on integration site datasets in columns and different genomic features in rows. For each comparison, integration site distributions are compared with distributions of randomly selected sites in the human genome.[4,35] Early integration site recovery methods involved use of cleavage by restriction enzymes, which are unevenly distributed in the human genome.[32,33] For this reason, random controls were matched based on proximity to restriction enzyme cleavage sites.[35] The INSPIIRED pipeline uses random cleavage by sonication, so purely random control sites are generated in silico and used in the analysis described here.

In the heatmaps, colored tiles indicate the intensity and direction of any departures from the distributions of random controls for each genomic feature in each integration site dataset. Three random sites are picked per integration site. The locations are then annotated using the hiAnnotator R package.

The coincidence of genomic feature "J" with each integration site and random control site is measured. The nonparametric method of estimating receiver operating characteristic (ROC) curve areas and their covariance structure of DeLong et al.[45] is used. Each integration site is compared in a pairwise fashion with random control sites, and a number is assigned indicating the relative rank of the integration site: 1 if the measurement of J is higher at the integration site than at a random control site, 0 if the measurement of J is lower at the integration site than at a random control site, and 0.5 if the measurement of J is equal for the two sites. All such values are calculated for a dataset of integration sites and averaged to obtain the overall ROC area for the feature measured (https://github.com/BushmanLab/hotROCs). This is equivalent to comparing the ranks of the sites with those of the controls. In older datasets with integration sites recovered by cleavage with restriction enzymes, matched random controls based on proximity to restriction enzyme cleavage sites were used. In that setting ROC curve areas were based on comparing each site only with its matched controls.[35]

An ROC area between 0 and 0.5 indicates the genomic feature occurs less frequently at or near integration sites than at or near random sites in the genome and is therefore disfavored. An ROC area between 0.5 and 1 indicates the genomic feature is enriched at integration sites. An ROC area of exactly 0.5 indicates that integration sites in the dataset are neither enriched nor depleted with respect to the feature of interest. The ROC area is converted to a color tile according to the colorimetric scale shown at the bottom of the heatmap.[35] In Figure 2, positive associations (enrichment compared with random) are shown as increasingly intense shades of blue, negative associations (depletion compared with random) as increasing intense shades of yellow, and no difference from random as black. Each tile represents a comparison of integration sites with the randomly sampled controls for one genomic feature (row) in one experimental dataset (column).

Note that we do not present the magnitude of effect in terms of the original units of measurement. We simply ask whether the average integration site has a higher rank for a given type of feature than its matched random control sites. The color indicates the average quantile of each integration site relative to its random controls. This removes skewing effects contributed by non-normal distributions of the data and also reduces the effect of data points with extreme values for a feature. Statistical tests are carried out to determine whether the ROC areas calculated are significantly different from one another or from 0.5 (indistinguishable from random controls; methods are further explained in Supplemental Materials and Methods).[35,46]

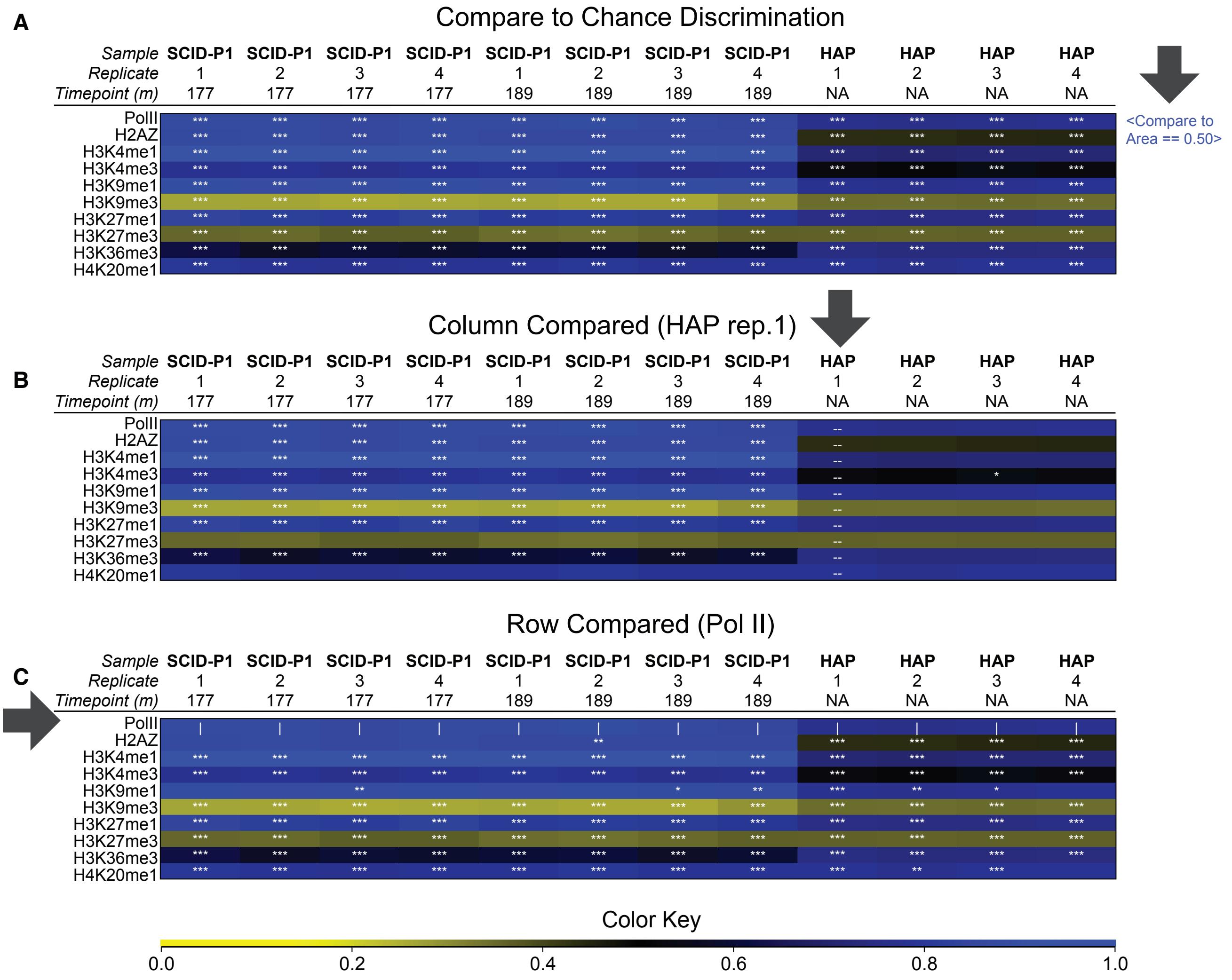

Figure 3. Probing Statistical Associations between Epigenetic Marks and Integration Site Density Using Interactive Heatmaps

Interactive heatmaps are available in Supplemental Information and Data S1. Labeling is as in Figure 2. Once heatmaps are loaded into an internet browser, clicking on different points on the image results in specific statistical tests, where results are summarized as asterisks on each tile of the heatmap (*p < 0.05; **p < 0.01; ***p < 0.001). The heavy black arrow in each panel indicates the selection by point and click. (A) Clicking on the text "Compare to area=0.05" yields statistical tests comparing the value for integration site data in each cell with random controls. (B) Comparison of outcome for each integration site dataset with the leftmost replicate of the data for lentiviral integration in HAP-1 cells (clicking on the leftmost HAP-1 column as indicated). All of the SCID-X1 gammaretroviral samples are different for each mark (indicated by the three asterisks [***]) except H3K27me3 and H4K20me1. (C) Comparison of distributions of integration sites relative to Pol II with the distribution of integration sites relative to other marks (click on Pol II). Seven out of eight are different, although for H2AZ in the gammaretroviral data, most show similar distributions (no asterisks).

The heatmap shown in Figure 2 compares the integration site distribution at two time points from a gene-corrected patient (patient 1 [P1]) with epigenetic marks mapped in CD133^{+} progenitor cells. Patient 1 (P1) was treated with an early gammaretroviral vector, used to deliver the missing IL2RG gene to treat SCID-X1.[47,48] The two samples were isolated from peripheral blood mononuclear cells (PBMCs) taken at 177 and 189.5 months after gene therapy. For comparison, another sample using a lentiviral-vector-infected human-derived HAP-1 cell line has been included to illustrate differences with the lentiviral integration pattern. Quadruplicate assays for each DNA sample are shown to illustrate reproducibility.

The distributions of integration sites datasets (Table S1) were compared with the distributions of 10 different epigenetic marks or bound DNA binding proteins.[49] Each of these was mapped by chromatin immunoprecipitation sequencing (ChIP-seq), in which each protein was covalently cross-linked to DNA, and bound DNA fragments were recovered by immunoprecipitation, sequenced, and then mapped to the human genome to identify relative density. Densities of mapped ChIP-seq annotations were compared with distributions of integration sites within 10 kb windows, and the collection of values was used to generate ROC areas.

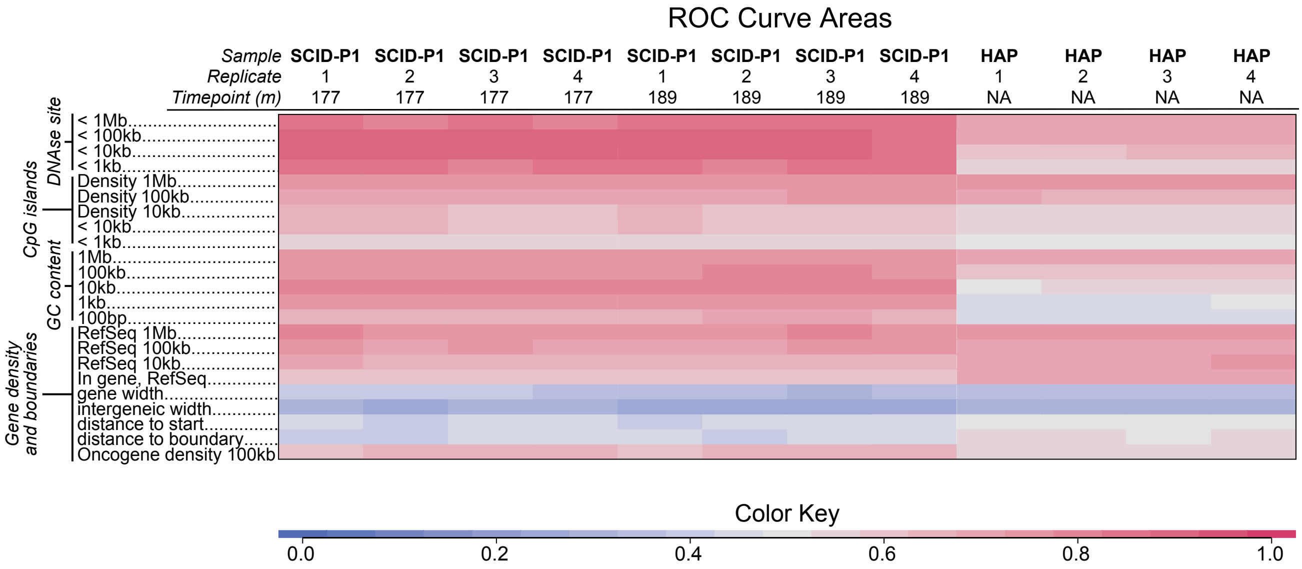

Figure 4. Heatmap Summarizing the Density of Integration Sites Relative to Genomic Features

Samples are shown in the columns, and features mapped onto the human genome are shown in rows. Associations are quantified using the ROC area method. The values of ROC areas are shown in the color key at the bottom. The numbers on the left indicate the lengths of genomic intervals used in comparisons with random controls. Oncogene density (bottom row) involves asking how frequently integration sites are found with 100 kb of transcription start sites for genes in the allOnco gene list (http://www.bushmanlab.org/links/genelists) compared with random controls.

For the gene therapy specimens made by infection of stem cells with a gammaretroviral vector (patient 1 [P1]), the distribution mostly favored marks associated with active transcription (H3K9me1, H3K4me1, H4K20me1, and H3K27me1). Integration was disfavored near marks associated with repressive chromatin (H3K27me3 and H3K9me3). However, the gammaretroviruses favor integration near transcription start sites,[10] and H2AZ and H3K4me3 were positively associated, and RNA polymerase II (Pol II) more strongly than for lentiviruses.

For the lentiviral infection in HAP1 cells,[50] integration is favored near sites of covalent modifications of histones associated with active transcription, including H3K9me1, H3K4me1, H4K20me1, H3K27me1, and H3K6me3. Lentiviral integration is favored within transcription units,[1,3,4] where the H3K36me3 mark is found, so integration was favored near this mark for lentiviral infection, but not gammaretroviral infection. Repressive chromatin marks were disfavored, including H3K27me3 and H3K9me3. Marks found near transcription start sites (H3K4me3 and H2AZ) were either slightly disfavored or neither favored nor disfavored.[51] These patterns parallel those seen previously for lentiviral vectors in diverse cell types.[4,8,11,52,53]

An added feature of these heatmaps is that they have been engineered to allow interactive statistical tests (Figure 3; interactive heatmaps are available in Supplemental Information and Data S1). Heatmaps are generated as scalable vector graphics (SVG), which can be opened in an Internet browser. Users can click on a row or column, and statistical results appear on the heatmap tiles documenting whether results in other rows or columns differ from the query. Users can also click on a button to the right of the maps to allow comparison of all tiles with the random control. Results of statistical comparisons are reported as asterisks on each tile. Some examples are shown in Figure 3, illustrating comparisons among random (Figure 3A), the leftmost HAP1 dataset (Figure 3B), or the Pol II ChIP-seq distribution (Figure 3C).

Figure 4 presents another form of the heatmap that queries the results of multiple additional features, including mapped DNase I cleavage site (which reports DNA accessibility), CpG islands (important in gene regulation), guanine/cytosine (GC) percentage, gene counts as documented in the refSeq dataset, and proximity to gene boundaries. For those features that are mapped in intervals (GC percentage over 1 Mb, 100 kb, 10 kb, and so on), it is often unknown a priori what width is the most relevant to the biological question at hand. Thus, for these features, results for a number of different interval sizes are shown.

All three datasets are compared over their four replicates against these features (Figure 4). High densities of DNase I hypersensitive sites and high densities of CpG islands are associated with favored integration for both lentiviral and gammaretroviral vectors (red coloration). High GC content is also favored, likely because high GC content is characteristic of gene-rich regions, although for lentiviruses, the preference switches to local high adenine/thymine (AT), possibly associated with binding of LEDGF/p75, the tethering cofactor, or wrapping of integration target site DNA on nucleosomes.[11] Paralleling the favored high GC content, direct measures of gene richness are also positively

A

Trial	GTSP	Replicates	Patient	Timepoint	CellType	TotalReads	InferredCells	UniqueSites	FragMethod	VCN	S.chao1	Gini	Shannon	UC50
SCIDn1	GTSP0518	4	pP1	m177	PBMC	549216	13803	1091	Shearing	0.4	1762	0.8505	4.4947	10
SCIDn1	GTSP0853	4	pP1	m189.5	PBMC	1008346	13756	975	Shearing	NA	1861	0.8642	4.2533	8

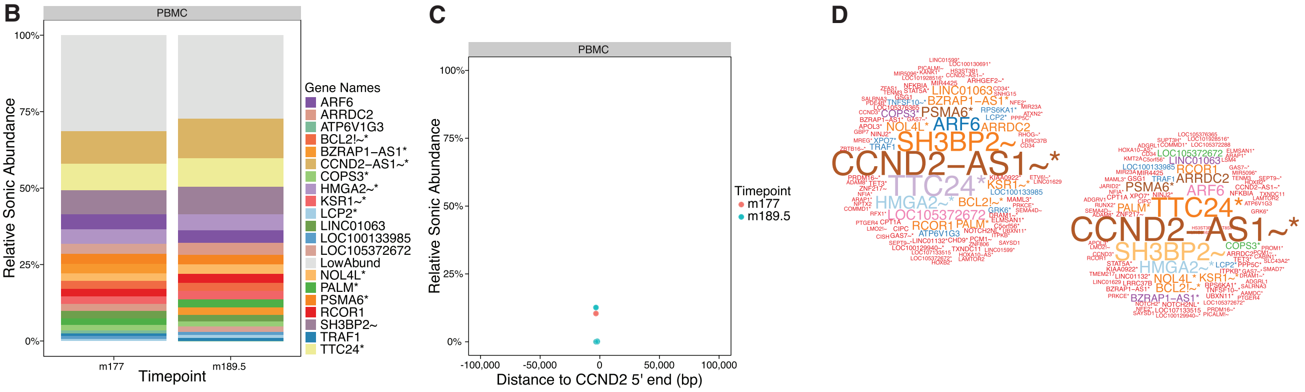

Figure 5. Excerpts from a Reproducible Report on SCID-X1 Patient 1

(A) Table summarizing sample metadata, including the trial, internal tracking number (GTSP), number of replicates, patient, time point queried, cell type, total number of sequence reads (TotalReads), inferred number of cells queried from SonicAbundance (InferredCells), the number of integration sites recovered after dereplication (UniqueSites), the method used to break the DNA (shearing in this case), the vector copy number if available (VCN) determined from qPCR, the minimum population size inferred from sharing among replicates (S.chao1), the asymmetry of clonal distribution (Gini), the diversity summarized as the Shannon index (Shannon), and the number of unique clones making up the top 50% of the sample abundance (UC50). (B) Stacked bar graph showing the most abundant clones, named after the nearest gene. Genes are annotated by whether the site is within a transcription unit (*), whether the site is within 50 kb of a cancer-related gene (~), or whether the site is associated with a gene strongly associated with human lymphoma (!). (C) Graph indicating the position of integration sites near CCND2, and their proportions as inferred by SonicAbundance. (D) Word bubbles summarizing the proportions of integration sites near each named gene. The size of the gene name in the word bubble is a function of the SonicAbundance of that site. Note that there is an antisense transcript upstream of the CCND2 transcription start site; thus, the integration site upstream is reported as CCND2-AS1 because it is within the DNA transcribed in the antisense transcript.

associated. Integration is disfavored relative to regions with long gene widths or long intergenic distances, because these are indicative of gene-sparse regions. The gammaretroviral vector sites are disfavored relative to long gene boundary distances because they favor integration near transcription start sites. Integration is favored within genes (as annotated by the refSeq dataset) for lentiviral vectors,[3] but only weakly favored for gammaretroviral vectors.

Thus, numerous relationships between integration site datasets and genomic features can be explored statistically using these interactive heatmaps.

Lists of Cancer-Associated Genes for Annotating Integration Site Distributions

A question of interest in many therapeutic applications centers on whether integration sites accumulate near the transcription start sites of cancer-related genes. A complication is that there are many ways of defining cancer-associated genes, and most such genes are important only in specific types of human cancers. For annotating gene therapy results, we have thus generated multiple lists of cancer-associated genes that can be queried as appropriate for integration site analysis (http://www.bushmanlab.org/links/genelists).

In one approach, we created a maximally comprehensive list (AllOnco) for use in first-pass screening based on the idea that we cannot predict what cancer-associated genes are most important in the novel clinical setting of human gene modification. The list incorporates known human cancer genes and human homologs of cancer genes in model organisms, and so includes 2,125 total genes, or roughly 8.5% of all human genes (assuming 25,000 total). Comparison with oncogene annotation is summarized using the heatmap format (Figure 4, bottom row), which scores the frequency of integration sites within 100 kb of cancer gene transcription start sites in integration sites versus random sites.

Reports on Integration Site Sample Sets for Tracking Outcome in Human Gene Therapy

An important application of integration site analysis is tracking outcome in human gene therapy. For this we have developed a standardized patient report template that rests on top of the INSPIIRED pipeline. Use of a reproducible report format allows tracking of datasets used in each study and version control of code (which are specified by dates). The report software takes in integration site and break point positional information, annotates the sites using hiAnnotator, and outputs targeted analyses of integration site distributions. An example of a patient report is provided in Data S2, showing two recent time points monitored for patient 1 (P1) treated for SCID-X1. Earlier time points were analyzed by 454 Roche pyrosequencing and were previously reported.[25]

Some excerpts from the report are presented in Figure 5. The software generates a summary table (Figure 5A) that reports the patient, time point, cell type, patient metadata, and summary statistics for each sample. Among these are the total numbers of reads, the number of cells inferred to have been sampled (sum of break points captured), and the number of unique integration sites after dereplication. Four statistics summarizing population structure are also calculated for each sample. The minimum population size is inferred from a Chao1 analysis with jackknife correction, which takes advantage of the four replicate analyses typically run for each sample.[34] Skewing in proportional abundance is calculated using the Gini index, where 0 indicates a perfectly even distribution of integration sites over the cells sampled and increases up to 1 with increasing oligoclonality. Diversity is calculated using the Shannon index, which summarizes both the number of different unique integration sites and the evenness of distribution of cells sampled (SonicAbundance) among integration sites.

Here, we introduce a new metric, called the UC50 (unique cell progenitors contributing the most to the expanded 50% of progeny cell clones). To generate the number, progenitor cell clones (reported as unique integration sites) are first ranked by the relative abundance of progeny cells using SonicAbundance (reported by linker ligation site data). The UC50 then reports the number of unique clones (integration sites) responsible for making up the top 50% of all cells sampled. Thus, if a single clone comprises more than 50% of the sample, the UC50 value will be 1. In contrast, for efficient lentiviral infections of cells in short-term tissue culture, the UC50 values can be in the thousands (data not shown).

Finally, where available, the vector copy number per cell (VCN), determined separately by qPCR, is added to allow assessment of the efficiency of gene marking in the cell population.

The relative abundance of integration sites in or near specific genes is summarized in several ways (Data S2). These include two types of stacked bar graphs (an example is shown in Figure 5B; both are displayed in Data S2). Bar graphs either display the number of cells observed (sonic breaks) that are associated with integration sites in the samples (scaled to the most abundant sample) or the proportional abundance of integration sites in each sample, where every sample is scaled to 100% (Figure 5B). Each high-abundance integration site is named by the closest gene, which is color coded in the key for each graph. Genes are annotated by whether the site is within a transcription unit (*), is within 50 kb of a cancer-related gene (~), or is associated with a gene strongly associated with human lymphoma (!). This last list consists of 38 human genes commonly involved in lymphoid cancers and includes most genes previously implicated in adverse events in human stem cell gene therapy, including LMO2, MDS/EVI1, and CCND2.[29–31]

Heatmaps provide another type of visualization (Data S2). These have the advantage over stacked bar graphs in that each integration site above a chosen abundance is given equal space, whereas in stacked bar graphs, rarer sites can be shown as thin bands that may be difficult to visualize. A third visualization is provided by line graphs (Data S2), which highlight the behavior of the most abundant clones over time. Another set of figures queries integration sites near genes of concern for adverse events in human gene therapy, including LMO2, CCND2, HMGA2, and MECOM.[29,37] In these visualizations, the distance from the transcription start site is shown on the x axis, proportional abundance is shown on the y axis, and the time point is color coded. By this means it can be seen that an integration site near CCND2 achieves >10% abundance near the CCND2 transcription start site (Figure 5C).

Handling integration sites that map to multiple locations in the human genome presents a particular challenge. It could be that an integration site authentically resides in a repeated sequence that is also in the 5′ region of a cancer-associated gene and potentially marking an adverse event. To accommodate this, reports include an account of the SonicAbundance of cells with integration sites in repeated sequences,[34] allowing tracking of possible sites of concern in multihits.

The reports end with word bubbles (Figure 5D), which provide a visual summary of genes near integration sites in expanded clones.[27] Names of genes that are nearest to the integration sites are used to construct the word bubble. Word size is scaled by the SonicAbundance measure for each integration site, and each gene name is marked with the same integration flags (*, ~, and !) used in earlier plots. By this means, the major clones in the sample are evident at a glance.

Outcome in the First SCID-X1 Gene Therapy Trial

Figure 5 shows excerpts from a reproducible report summarizing monitoring of patient 1 (P1) from the first trial to treat SCID-X1. Results are summarized for PBMCs from two time points, 177 and 189.5 months after gene therapy. Half a million to a million reads were collected for each sample, allowing investigation of 13,000 cells associated with about 1,000 integration sites. UC50 values for the two time points are 10 and 8, indicating the presence of expanded cell clones.

In early studies based on 454/Roche pyrosequencing, the subject was found to have an expanded clone with an integration site near CCND2 (6% of all reads), a gene for which a nearby integration event was associated with an adverse event in another SCID-X1 gene-corrected patient.[25,29] Thus, it was of interest to monitor the behavior of the clone in this patient over time. Analyses using Illumina paired-end sequencing are summarized in Figure 5, which shows that the CCND2 clone has slightly expanded in abundance (nonparametric comparison of replicate medians yields $p = 0.029$ when compared by relative abundances and $p = 0.057$ when compared by absolute abundances judged by SonicAbundance). The integration site is 3,241 nt upstream of the CCND2 transcriptional start site. Thus, longitudinal tracking reveals a stable expanded clone in this subject.

DISCUSSION

Here, we describe a collection of tools for the analysis and visualization of integration site distributions. This tool set takes advantage of the INSPIIRED pipeline described in the accompanying paper.[9] Integration sites are mapped to the human genome, the abundances of their host cells tabulated using the SonicAbundance method, and results stored in a database, allowing flexible downstream analysis. Integration site distributions can then be compared with genomic features and sites of epigenetic modification as ROC areas. These data are summarized as interactive heatmaps, allowing comparison with random distributions, other integration site datasets, or genomic annotation by simply clicking on a row or column, which outputs comprehensive statistical tests.

For applications to human gene therapy, interest often focuses on longitudinal behavior of expanded clones and proximity of integration sites to cancer-associated genes. A standardized report format was developed, allowing interactive comparison among patient datasets and querying multiple aspects of longitudinal behavior. For this, we introduce the UC50 metric, which is generated by ranking progenitors (integration sites) from most to fewest daughter cells produced (linker ligation sites) and counting the number of progenitors contributing to the top 50% of the distribution. Thus, clonal expansion yields low UC50 numbers and highly polyclonal samples high UC50 numbers. These tools were used to query recent clonal behavior in a patient from the first SCID-X1 gene therapy trial.[29,47,48] The patient studied has an expanded clone with an integration site near the proto-oncogene CCND2. The analysis of integration sites from month 177 to month 189.5 post-treatment revealed stability of this clone, with possible slow expansion. This analysis illustrates how the tools described here can be applied to monitor outcomes in gene therapy.

MATERIALS AND METHODS

Human Subjects

As in Cavazzana-Calvo et al.[47] and Hacein-Bey-Abina et al.,[48] patient 1 (P1) fulfilled the eligibility requirements for first ex vivo γc gene therapy trial (1999–2002) at age 11 months. P1 was diagnosed with SCID-X1 based on his blood lymphocyte phenotype, revealing a tail-less γc receptor expressed at the membrane (R289 X). Marrow was harvested and subjected to $CD34^+$ cell separation, obtaining 9.8×10^6 $CD34^+$ cells per kilogram of body weight. Harvested cells were then exposed to MFG γc vector-containing supernatant daily for 3 days. P1 was then infused with the treated $CD34^+$ cells (19×10^6 cells/kg) without prior chemoablation.

Integration Site Analysis

As explained in the companion paper,[9] integration sites are identified by sequencing the LTR-host junctions from genomic DNA after linker-mediated PCR amplification. Genomic DNA is randomly sheared by ultrasonication, after which linkers are ligated to the repaired DNA for amplification. Nested PCR is used to amplify the LTR-host DNA junctions by priming from the viral LTR and the linkers, appending the sequences needed for sequencing. Samples are sequenced using the Illumina paired-end platform, and the output sequencing files are processed by intSiteCaller to yield integration site positions on a host draft genome. Integration site data and ChIP-seq data were mapped onto the hg18 genome draft, to match the original draft genome used for analysis of the ChIP-seq data. As in Berry et al.,[35] receiver operating characteristic (ROC) areas are used to compare integration sites with random control sites.

Pipeline Utilization

INSPIIRED is distributed online as a downloadable virtual machine executable on the Windows, Mac, and Linux operating systems, as well as a GitHub source code repository supported by a Conda software environment (see https://github.com/BushmanLab/INSPIIRED, which also includes detailed instructions for use and test datasets).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Materials and Methods, one figure, one table, and two data files and can be found with this article online at http://dx.doi.org/10.1016/j.omtm.2016.11.003.

AUTHOR CONTRIBUTIONS

E.S., C.N., C.C.B., E.S., M.C., and F.D.B. designed the study; E.S., C.N., C.C.B., E.S., M.C., F.M., M.J.D., P.B., S.H.-B.-A, L.C., and F.D.B. carried out the study; E.S., C.N., C.C.B., E.S., Y.W., A.D., N.M., A.B., K.B., J.K.E., M.C., F.M., S.R., M.J.D., P.B., S.H.-B.-A., L.C., and F.D.B. analyzed the data.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

We are grateful to members of the F.D.B. laboratory for help and suggestions. All authors were supported by grants AI 052845, AI 104400, AI 082020, AI 045008, AI 117950, and HL 113252 and ERC Regenerative Therapy grant 269037 from the European Research Council, and an award from the French ANRS. We also acknowledge support from the Penn Center for AIDS Research (grant P30 AI 045008) and the PennCHOP Microbiome Program.

REFERENCES

1. Bushman, F.D. (2001). Lateral DNA Transfer: Mechanisms and Consequences (Cold Spring Harbor Laboratory Press).

2. Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. (2002). Mobile DNA II (Washington, D.C.: American Society for Microbiology Press).

3. Schröder, 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.

4. Mitchell, R.S., Beitzel, B.F., Schroder, A.R., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*, E234.

5. Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). Retroviruses (Cold Spring Harbor Laboratory Press).

6. Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F., et al. (2014). HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science *345*, 179–183.

7. Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y., Larsen, B.B., Styrchak, S., Huang, H.C., Edlefsen, P.T., Mullins, J.I., and Frenkel, L.M. (2014). HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science *345*, 570–573.

8. Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., Hahn, B.H., Czartoski, J.L., McElrath, M.J., Lehmann, C., et al. (2015). HIV-1 integration landscape during latent and active infection. Cell *160*, 420–432.

9. Sherman, E., Nobles, C., Berry, C., Six, E., Wu, Y., Dryga, A., Malani, N., Male, F., Reddy, S., Bailey, A., et al. (2017). INSPIIRED: A pipeline for qualtitative analysis of sites of new DNA integration in cellular genomes. Mol Ther Methods Clin Dev. *4*, 39–49.

10. 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.

11. 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.

12. Biffi, A., Bartolomae, C.C., Cesana, D., Cartier, N., Aubourg, P., Ranzani, M., Cesani, M., Benedicenti, F., Plati, T., Rubagotti, E., et al. (2011). Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. Blood *117*, 5332–5339.

13. Calabria, A., Leo, S., Benedicenti, F., Cesana, D., Spinozzi, G., Orsini, M., Merella, S., Stupka, E., Zanetti, G., and Montini, E. (2014). VISPA: a computational pipeline for the identification and analysis of genomic vector integration sites. Genome Med. *6*, 67.

14. Hocum, J.D., Battrell, L.R., Maynard, R., Adair, J.E., Beard, B.C., Rawlings, D.J., Kiem, H.P., Miller, D.G., and Trobridge, G.D. (2015). VISA–Vector Integration Site Analysis server: a web-based server to rapidly identify retroviral integration sites from next-generation sequencing. BMC Bioinformatics *16*, 212.

15. Rae, D.T., Collins, C.P., Hocum, J.D., Browning, D.L., and Trobridge, G.D. (2015). Modified genomic sequencing PCR using the MiSeq platform to identify retroviral integration sites. Hum. Gene Ther. Methods *26*, 221–227.

16. LaFave, M.C., Varshney, G.K., and Burgess, S.M. (2015). GeIST: a pipeline for mapping integrated DNA elements. Bioinformatics *31*, 3219–3221.

17. Jiang, H., and Wong, W.H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics *24*, 2395–2396.

18. Peters, B., Dirscherl, S., Dantzer, J., Nowacki, J., Cross, S., Li, X., Cornetta, K., Dinauer, M.C., and Mooney, S.D. (2008). Automated analysis of viral integration sites in gene therapy research using the SeqMap web resource. Gene Ther. *15*, 1294–1298.

19. Hawkins, T.B., Dantzer, J., Peters, B., Dinauer, M., Mockaitis, K., Mooney, S., and Cornetta, K. (2011). Identifying viral integration sites using SeqMap 2.0. Bioinformatics *27*, 720–722.

20. Hacein-Bey Abina, S., Gaspar, H.B., Blondeau, J., Caccavelli, L., Charrier, S., Buckland, K., Picard, C., Six, E., Himoudi, N., Gilmour, K., et al. (2015). Outcomes following gene therapy in patients with severe Wiskott-Aldrich syndrome. JAMA *313*, 1550–1563.

21. Hacein-Bey-Abina, S., Pai, S.Y., Gaspar, H.B., Armant, M., Berry, C.C., Blanche, S., Bleesing, J., Blondeau, J., de Boer, H., Buckland, K.F., et al. (2014). A modified γ-retrovirus vector for X-linked severe combined immunodeficiency. N. Engl. J. Med. *371*, 1407–1417.

22. Ott, M.G., Schmidt, M., Schwarzwaelder, K., Stein, S., Siler, U., Koehl, U., Glimm, H., Kühlcke, 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.

23. Kustikova, O.S., Baum, C., and Fehse, B. (2008). Retroviral integration site analysis in hematopoietic stem cells. Methods Mol. Biol. *430*, 255–267.

24. Biasco, L., Baricordi, C., and Aiuti, A. (2012). Retroviral integrations in gene therapy trials. Mol. Ther. *20*, 709–716.

25. 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.

26. Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G.P., Berry, C.C., Martinache, C., Rieux-Laucat, F., Latour, S., Belohradsky, B.H., et al. (2010). Efficacy of gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. *363*, 355–364.

27. Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M.P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M.C., et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. Science *341*, 1233151.

28. Baum, C. (2007). Insertional mutagenesis in gene therapy and stem cell biology. Curr. Opin. Hematol. *14*, 337–342.

29. 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.

30. 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.

31. Braun, C.J., Boztug, K., Paruzynski, A., Witzel, M., Schwarzer, A., Rothe, M., Modlich, U., Beier, R., Göhring, G., Steinemann, D., et al. (2014). Gene therapy for Wiskott-Aldrich syndrome–long-term efficacy and genotoxicity. Sci. Transl. Med. *6*, 227ra33.

32. 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.

33. 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.

34. Berry, C.C., Gillet, N.A., Melamed, A., Gormley, N., Bangham, C.R., and Bushman, F.D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics *28*, 755–762.

35. 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.

36. Berry, C.C., Ocwieja, K.E., Malani, N., and Bushman, F.D. (2014). Comparing DNA integration site clusters with scan statistics. Bioinformatics *30*, 1493–1500.

37. 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 β-thalassaemia. Nature *467*, 318–322.

38. Brady, T., Roth, S.L., Malani, N., Wang, G.P., Berry, C.C., Leboulch, P., Hacein-Bey-Abina, S., Cavazzana-Calvo, M., Papapetrou, E.P., Sadelain, M., et al. (2011). A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. Nucleic Acids Res. *39*, e72.

39. Wang, G.P., Levine, B.L., Binder, G.K., Berry, C.C., Malani, N., McGarrity, G., Tebas, P., June, C.H., and Bushman, F.D. (2009). Analysis of lentiviral vector integration in HIV+ study subjects receiving autologous infusions of gene modified CD4+ T cells. Mol. Ther. *17*, 844–850.

40. Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., Baldoli, C., Martino, S., Calabria, A., Canale, S., et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. Science *341*, 1233158.

41. Cassani, B., Montini, E., Maruggi, G., Ambrosi, A., Mirolo, M., Selleri, S., Biral, E., Frugnoli, I., Hernandez-Trujillo, V., Di Serio, C., et al. (2009). Integration of retroviral vectors induces minor changes in the transcriptional activity of T cells from ADA-SCID patients treated with gene therapy. Blood *114*, 3546–3556.

42. Gillet, N.A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., Berry, C., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2011). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood *117*, 3113–3122.

43. Olszko, M.E., Adair, J.E., Linde, I., Rae, D.T., Trobridge, P., Hocum, J.D., Rawlings, D.J., Kiem, H.P., and Trobridge, G.D. (2015). Foamy viral vector integration sites

in SCID-repopulating cells after MGMTP140K-mediated in vivo selection. Gene Ther. *22*, 591–595.

44. LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. *42*, 4257–4269.

45. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics *44*, 837–845.

46. Ocwieja, K.E., Brady, T.L., Ronen, K., Huegel, A., Roth, S.L., Schaller, T., James, L.C., Towers, G.J., Young, J.A., Chanda, S.K., et al. (2011). HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. PLoS Pathog. *7*, e1001313.

47. 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.

48. 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.

49. Raney, B.J., Dreszer, T.R., Barber, G.P., Clawson, H., Fujita, P.A., Wang, T., Nguyen, N., Paten, B., Zweig, A.S., Karolchik, D., and Kent, W.J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics *30*, 1003–1005.

50. Petersen, J., Drake, M.J., Bruce, E.A., Riblett, A.M., Didigu, C.A., Wilen, C.B., Malani, N., Male, F., Lee, F.H., Bushman, F.D., et al. (2014). The major cellular sterol regulatory pathway is required for Andes virus infection. PLoS Pathog. *10*, e1003911.

51. 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.

52. 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.

53. Marshall, H.M., Ronen, K., Berry, C., Llano, M., Sutherland, H., Saenz, D., Bickmore, W., Poeschla, E., and Bushman, F.D. (2007). Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS ONE *2*, e1340.