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. Bushman2

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 of the expanded clone. This allows for accurate quantification and analyzing such data, but optimal methods 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...
published statistical tools for analysis of such data and applied these tools to track several gene therapy trials. Other groups have also used related methods.

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-XI (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.

RESULTS

The INSPIRIED Pipeline
The INSPIRIED 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). 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 INSPIRIED 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. All downstream analyses are carried out using genomic locations and sonic break points. The IntSiteDB schema is shown in Figure S1.
These maps were introduced in Berry et al.,35 which presents more genomic features in rows. For each comparison, integration site dis-

formations on integration site datasets in columns and different sites.35 The INSPIIRED pipeline uses random cleavage by sonication, controls were matched based on proximity to restriction enzyme cleavage distributed in the human genome.32,33 For this reason, random con-

rivied use of cleavage by restriction enzymes, which are unevenly involved in the human genome.4,35 Early integration site recovery methods

site distributions and genomic annotations (Figures 2, 3, and 4). We use heatmaps to summarize the relationships between integration sites 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 comb-

ined 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 separa-

tion 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 INSPIRED 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 esti-

mating 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 data-

set 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
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. 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. 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, which each protein was covalently cross-linked to DNA, and bound DNA fragments were recovered by immunoprecipitation, sequencing, 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.

The heatmap 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. 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.
For the gene therapy specimens made by infection of stem cells with a gammaretroviral vector (patient 1 [P1]), the distribution most 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, 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, integration is favored near this mark for lentiviral infection, but not 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, and H2AZ and H3K4me3 were positively associated, and RNA polymerase II (Pol II) more strongly than for lentiviruses.

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. Paralleling the favored high GC content, direct measures of gene richness are also positively correlated with GC content.
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 integra-
tion 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 incor-
porates 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). Compari-
son with oncogene annotation is summarized using the heatmap
format (Figure 4, bottom row), which scores the frequency of integra-
tion sites within 100 kb of cancer gene transcription start sites in inte-
gration 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 stan-
dardized patient report template that rests on top of the INSPIIRED
pipeline. Use of a reproducible report format allows tracking of data-
sets used in each study and version control of code (which are spec-
ified 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

| Trial | GTSP | Replicates | Patient | Timepoint | CellType | TotalReads | InferredCells | UniqueSites | FragMethod | VCN | S.chao1 | Gini | Shannon | UC50 |
|-------|------|------------|---------|-----------|----------|------------|---------------|-------------|------------|-----|---------|------|---------|------|
| ScION1 | GTSP0518 | 4 | p1p1 | m189.5 | PBMC | 1008346 | 13756 | 975 | Shearing | NA | 1961 | 0.8805 | 4.4684 | 10 |
| ScION1 | GTSP0653 | 4 | p1p1 | m189.5 | PBMC | 1008346 | 13756 | 975 | Shearing | NA | 1961 | 0.8805 | 4.4684 | 10 |

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.
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.22 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 site data. The UC50 values for the two time points are 10 and 8, indicating the presence of expanded cell clones.

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-correction 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 INSPIRED pipeline described in the accompanying paper. Integration sites are sequenced using the Illumina paired-end platform, and the output linkers, appending the sequences needed for sequencing. Samples LTR-host DNA junctions by priming from the viral LTR and the positions on a host draft genome. Integration site data and ChIP-seq in a patient from the UC50 numbers. These tools were used to query recent clonal behavior sion yields low UC50 numbers and highly polyclonal samples high contributing to the top 50% of the distribution. Thus, clonal expansion 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. 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. and Hacein-Bey-Abina et al., 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 × 10⁶ 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 × 10⁶ cells/kg) without prior chemoablation.

Integration Site Analysis
As explained in the companion paper, 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 re-paired 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., receiver operating characteristic (ROC) areas are used to compare integration sites with random control sites.

Pipeline Utilization
INSPIRED 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/INSPIRED, 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. Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, I.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.
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., Hiege, 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., Bouneau, 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.R., 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, 828–837.

52. Cuiffo, L., 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.
Supplemental Information

INSPIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions

Charles C. Berry, Christopher Nobles, Emmanuelle Six, Yinghua Wu, Nirav Malani, Eric Sherman, Anatoly Dryga, John K. Everett, Frances Male, Aubrey Bailey, Kyle Bittinger, Mary J. Drake, Laure Caccavelli, Paul Bates, Salima Hacein-Bey-Abina, Marina Cavazzana, and Frederic D. Bushman
Additional Methods. Navigating the Genomic and Epigenetic Heatmaps

General Description
Heatmaps are used to summarize the relationships between integration site distributions and various genomic or epigenetic features. Initially introduced in 2006 [1], these heatmaps are highly customizable. New features can be downloaded from the UCSC Genome Bioinformatics Group’s website (genome.ucsc.edu). This short guide is intended to give a brief overview of how to work with and interpret the heatmaps. More detailed information can be found in Ocwieja, K. et al. 2011 [2] (Supplementary information).

1. Opening the heatmap file
The output of the genomicHeatmapMaker and EpigeneticHeatmapMaker programs is a directory (or folder) which contains a series of “.svg” or scalable vector graphic files. These files can be opened with any current web browser (Chrome, Safari, Firefox, Internet Explorer, …). When initially opening the heatmap, find the file named “main.svg” and open it in your favorite browser.

Within the browser, the heatmap should be displayed with the title “ROC Curve Areas” at the top, and a color key at the bottom (default color scale from blue to red). Columns summarize integration site data sets (such as a single patient) while either genomic or epigenetic features are summarized in the rows. This is indicated by the name of the integration data sets (HAP-1, pP1, …) above each column and the name of each feature (CpG_density.1M, H3K4me1, …) next to each row on the left. The colored tiles indicate the intensity and direction of any departures from the distribution of random controls for each feature in each integration site data set.

2. Navigating statistical tests
The heatmap is an interactive figure. There are several areas of the heatmap that can be “clicked” or selected to yield statistical comparisons. By selecting a column name (such as a single patient), all integration site data sets are compared to the selected data set on for each feature independently. Likewise, selecting a row applies the same type of comparison between features. Selected columns or rows are used as controls and are marked by either “--” or “|”, respectively. To return to the plain heatmap, a link is available to the top right of the chart which will reset the comparison, labeled “Show Plain Heatmap”. Lastly, a link is available to the top right of the heatmap for comparing each tile to random controls, labeled “Compare to Area == 0.50”.

Comparisons between integration site data sets, features, and random controls utilize the Wald-test statistic and are referred to a Chi Square distribution to obtain p-values, as described in Brady, T. et al. 2009 [3]. The Wald-test statistic is constructed from the variance-covariance matrix of the relative ranks of the integration sites. P-value designation is as follows:

*=p<0.05, **=p<0.01, ***=p<0.001
References:

1. Berry C, Hannenhalli S, Leipzig J, Bushman FD: Selection of target sites for mobile DNA integration in the human genome. *PLoS computational biology* 2006, 2(11):e157.

2. Ocwieja KE, Brady TL, Ronen K, Huegel A, Roth SL, Schaller T, James LC, Towers GJ, Young JA, Chanda SK et al: HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* 2011, 7(3):e1001313.

3. Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, Bushman FD: Integration target site selection by a resurrected human endogenous retrovirus. *Genes & development* 2009, 23(5):633-642.
Additional Figure 1
Database Schema

- **samples**
  - sampleID INT(11)
  - sampleName VARCHAR(255)
  - refGenome VARCHAR(10)
  - gender CHAR(1)
  - miseqid VARCHAR(255)
  - comment VARCHAR(255)

- **multihitpositions**
  - multihitID INT(11)
  - sampleID INT(11)
  - position INT(11)
  - chr VARCHAR(255)
  - strand CHAR(1)

- **multihitlengths**
  - multihitID INT(11)
  - sampleID INT(11)
  - length INT(11)
  - count INT(11)

- **sites**
  - siteID INT(11)
  - sampleID INT(11)
  - position INT(11)
  - chr VARCHAR(255)
  - strand CHAR(1)

- **pcrbreakpoints**
  - siteID INT(11)
  - breakpoint INT(11)
  - count INT(11)
Table S1. List of DNA sequence data in manuscript

| Sample Name       | CellType | Time Point | Vector         | Origin                  | Accession Number |
|-------------------|----------|------------|----------------|-------------------------|------------------|
| SCID-X1 Patient P1 | PBMC     | Month 177  | MFG-B2 gamma-c vector | This publication     | SRR3882947       |
| SCID-X1 Patient P1 | PBMC     | Month 189.5| MFG-B2 gamma-c vector | This publication     | SRR3882947       |
| HAP-1 Library 4   | HAP-1    | Day 1      | LentiET vector   | Petersen, J., PLoS Path. 2014 | SRR3882945       |

(Carette, J., Nature 2011)