Molecular Therapy
Methods & Clinical Development
Original Article

INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes

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

[1]Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-6076, USA; [2]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA 92093, USA; [3]Imagine Institute, Paris Descartes-Sorbonne Paris Cité University, 75014 Paris, France; [4]Laboratory of Human Lymphohematopoiesis, INSERM 24, 75014 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

Integration of new DNA into cellular genomes mediates replication of retroviruses and transposons; integration reactions have also been adapted for use in human gene therapy. Tracking the distributions of integration sites is important to characterize populations of transduced cells and to monitor potential outgrow of pathogenic cell clones. Here, we describe a pipeline for quantitative analysis of integration site distributions named INSPIIRED (integration site pipeline for paired-end reads). We describe optimized biochemical steps for site isolation using Illumina paired-end sequencing, including new technology for suppressing recovery of unwanted contaminants, then software for alignment, quality control, and management of integration site sequences. During library preparation, DNAs are broken by sonication, so that after ligation-mediated PCR the number of ligation junction sites can be used to infer abundance of gene-modified cells. We generated integration sites of known positions in silico, and we describe optimization of sample processing parameters refined by comparison to truth. We also present a novel graph-theory-based method for quantifying integration sites in repeated sequences, and we characterize the consequences using synthetic and experimental data. In an accompanying paper, we describe an additional set of statistical tools for data analysis and visualization. Software is available at https://github.com/BushmanLab/INSPIIRED.

INTRODUCTION

Integration of new DNA is important in studies in many fields, including retroviral and transposon replication,[1–4] HIV latency,[5–7] and human gene therapy.[8–13] Distributions of integration sites are not random in the host cell genome but differ among different integrating elements.[1–3,14,15] For several cases, tethering of integration complexes to cellular proteins has been shown to influence integration target site selection.[1–3,16–19] Genomic alterations resulting from integration can contribute to preferential proliferation or survival of the modified cells. Examples include insertional activation by retroviruses in animal models,[1,4] outgrowth of cells in HIV latency,[5–7] accumulation of endogenous retroviruses evolutionarily in metazoan genomes,[1,2,20] and outgrowth of specific cell clones during human gene therapy.[21–28] Often, it is useful to track the behavior of cells harboring newly integrated DNA longitudinally using next-generation sequencing.

Previously, we and others have carried out sequence-based surveys of integration site distributions, using first Sanger sequencing, then 454/Roche pyrosequencing, and today Illumina sequencing.[6,7,9,11–13,15,29–34] The Illumina platform has the advantages of allowing paired-end sequencing and providing larger data volumes. Several reports have described methods for analysis of these data.[9,29,35–48] However, to date, none have taken full advantage of all types of paired reads, dealt comprehensively with integration in repeated sequences, or provided a statistical framework for quantitative inference of cell abundances based on integration site data.

Here we adapt statistical approaches reported in three previous publications to management of Illumina paired-end data.[49–51] We first describe optimized biochemical methods for integration site isolation, which achieve the critical criteria of suppressing PCR contamination between samples while sampling randomly from the pool of integrated DNAs. We then describe methods for alignment, data management, and quantification of cell clones based on integration site data. The pipeline accommodates analysis of integration in both single-copy and repeated sequences. We generated synthetic integration sites corresponding to known locations on the human genome and used them in tests to optimize performance of our pipeline, including quantifying the influence of error in sequence determination. Performance was then tested over several datasets ranging from experimental infections to human gene therapy samples, allowing analysis

Received 18 August 2016; accepted 15 November 2016;
http://dx.doi.org/10.1016/j.omtm.2016.11.002.

[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, 426 Johnson Pavilion, Philadelphia, PA 19104-6076, USA.
E-mail: bushman@mail.med.upenn.edu

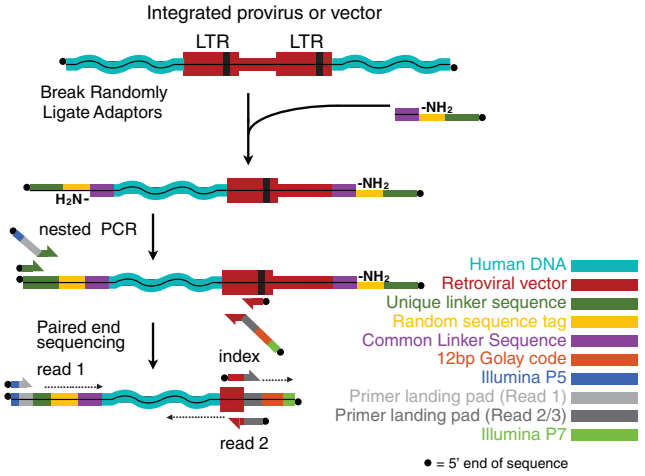

Figure 1. Diagram of the Biochemical Method for Isolating and Sequencing Sites of New DNA Integration

Genomic DNA containing an integrated retrovirus or retroviral vector (top) is sheared and DNA linkers are ligated onto the resulting ends (middle). The molecules are then subjected to two rounds of PCR amplification and then Illumina paired-end sequencing (bottom). The color code for sequence elements is summarized on the right. Black spheres indicate DNA 5′ ends. -NH_2 indicates an amino-modifier group, which prevents polymerase extension from the modified 3′ end.

of the influence of repeated sequences on site capture. In our accompanying paper in this issue of *Molecular Therapy: Methods & Clinical Development*,[52] we describe a suite of analytical tools that draws on the data products described here.

RESULTS

Biochemical Methods for Determining Sequences of Integration Acceptor Sites

Biochemical methods for recovering integration sites from DNA samples are diagrammed in Figure 1, and a detailed protocol is available in the Supplemental Materials and Methods and Tables S1–S8. Initially, isolated genomic DNA containing integration sites is randomly sheared by sonication. DNA linkers are then ligated to the sheared DNA ends. These DNA fragments are used as templates for PCR using one primer complementary to the linker and a second complementary to the end of the integrated DNA. In retroviruses and retroviral vectors, the ends of the integrated DNA correspond to the long terminal repeats (LTRs). Two rounds of PCR with nested primers are used to maximize specificity and recovery of sites from samples with small numbers of proviruses. Illumina sequencing adapters are attached to the DNA primers used for the second round of PCR, so that the PCR products generated contain the terminal sequences needed for sequence analysis on the MiSeq or HiSeq platforms.

The LTR sequences of an integrated provirus or retroviral vector are duplicated at each end of the integrated element—as a result, PCR using a primer complementary to the LTR results in amplification of two DNA products. One contains the desired flanking host

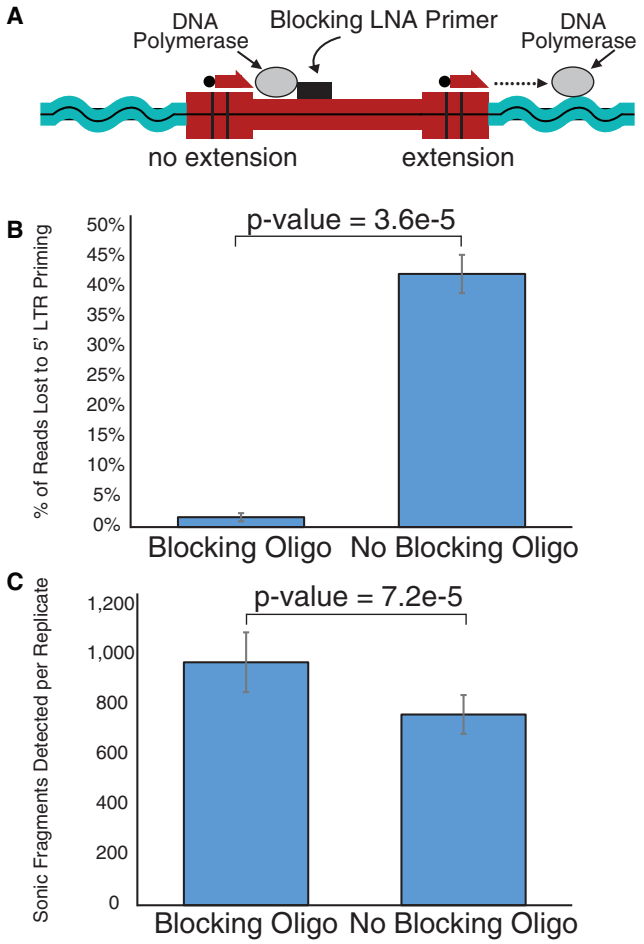

Figure 2. Use of a Locked Oligonucleotide to Block Recovery of the Internal Fragment to Improve Integration Site Yield

(A) Diagram of the method. The BNA-containing blocking oligonucleotide is shown in black; DNA polymerase is shown in gray. Other markings are as in Figure 1. (B) Quantification of recovery of the internal fragment. Twelve replicates were compared for each sample. (C) Increase in yield as a result of use of the locked oligonucleotide blocking primer. Error bars are SD.

DNA, and the second contains an unwanted internal sequence (Figure 2A). We thus developed a blocking oligonucleotide to reduce polymerase extension from the internal fragment. To increase affinity, blocking oligonucleotides were synthesized with multiple bases containing a bridging ring between the 2′ and 4′ positions.[53–55] The blocking oligonucleotide terminates with a 3′ amino modification to inhibit polymerase extension from the blocking oligonucleotide itself (Table S9).

In experiments comparing results with and without the blocking oligonucleotides (Figures 2B and 2C), inclusion of the blocking oligo reduced capture of the internal fragment from 42% to 1.6% of sequence reads and increased the average sampling of cellular genomes from 765 cells per replicate to 975 cells per replicate (as measured by SonicAbundance; described below).

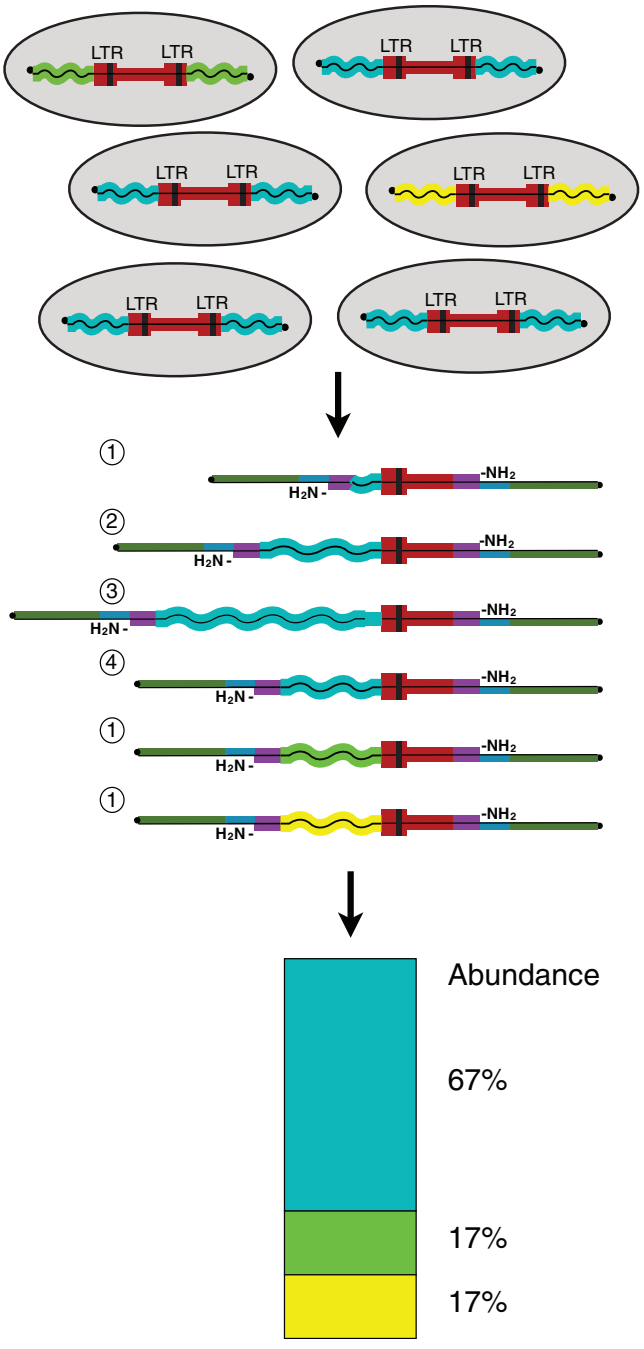

Figure 3. Estimating Abundance Using the SonicAbundance Method
Cells harboring integrated vectors are shown at the top. One cell clone has expanded to comprise 4/6 cells (flanking DNA colored cyan). DNA is then purified and cleaved and linkers are ligated. Note that the cyan expanded clone is present as four distinct fragment lengths. A stacked bar graph (bottom) summarizes the differences seen based on summing the abundance of different length fragments.

Given that multiple samples are commonly worked up simultaneously, and batches of samples may be analyzed frequently, PCR contamination between samples can be a severe problem. To suppress PCR contamination, each DNA sample is analyzed using 1 of 96 unique DNA linkers, which are paired with unique complementary PCR primers. Thus, any molecule moving between tubes would bear the wrong linker and thus would not be a substrate for PCR amplification.

Each sample is also given a unique 12-nucleotide (nt) error-correcting DNA barcode.[56–58] The combination of specific linker and barcode is rotated for each batch of samples processed, and correct pairing between bar code and linker sequences is required during quality filtering of output sequences (below). For all batches, negative controls are included, which are human DNA specimens lacking integrated vectors. Using these precautions, contamination due to PCR cross-over is rare or eliminated, as indicated by a consistent lack of recovery of integration sites from genomic DNA-only controls that lack integrated vectors.

To further mark each unique integration site sequence, each linker is synthesized with a random sequence of 12 nucleotides. Thus, linker ligation attaches a unique "primer ID" to each molecule prior to PCR.[59] These tags provide a potential means of abundance estimation by counting primer IDs, but in practice, this is complicated by PCR recombination (unpublished data). Thus, the main use in our pipeline is tracking possible contamination due to PCR cross-over between replicates by tracking primer IDs.

The SonicAbundance Method

We use the SonicAbundance method to infer the abundance of cell clones from integration site data (Figure 3).[50] Simply counting the number of sequence reads per integration site is known to yield distorted abundance estimates[32,60]—for example, shorter molecules amplify more efficiently than longer ones. The SonicAbundance method takes advantage of marks introduced into DNA molecules by sonication and linker ligation prior to the PCR amplification steps.

In a DNA sample from cells containing integration sites, an integration site from an expanded cell clone will be found in many cell genomes (Figure 3, top). Fragmentation by sonication followed by linker ligation results in many linkers joined near the integrated provirus from the expanded clone (Figure 3, middle). PCR amplification and paired-end sequencing results in recovery of many different sites of linker ligation near the unique integration site in the expanded clone. Sites of linker ligation are recovered in read 1, and LTR-host junctions are recovered in read 2 (Figure 1). The number of these linker positions is tabulated, providing an abundance score (Figure 3, bottom). For statistical analysis, the estimated abundance needs to be corrected to account for the frequency of identical linker ligation positions generated independently, which occurs with increasing frequency as the numbers of linker positions increases per integration site.[50] Numbers of linker ligation sites are recorded along with integration site positions and uploaded into our IntSiteDB database for analysis.

Processing and Aligning Integration Site Sequence Data

INSPIIRED begins by parsing raw Illumina output files (FASTQ format) using both index and linker sequences. Indexes are based

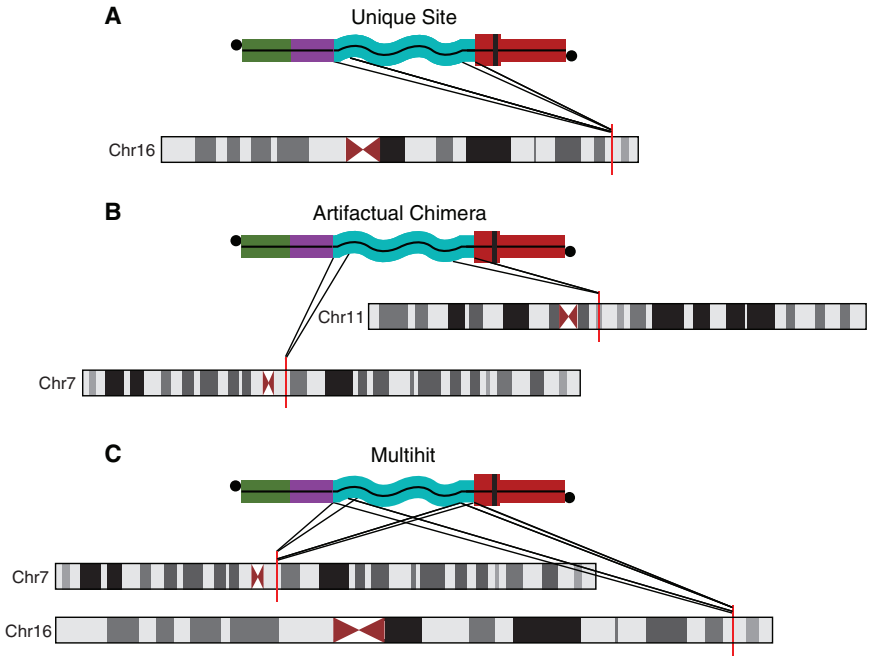

Figure 4. Interpretation of Paired Read Data
(A) Unique integration site. In this case, the two reads are within a short distance of one another on the chromosome and are correctly oriented on opposite DNA strands. (B) An artifactual chimera. In this case, the two reads are on two distinct chromosomes, or found implausibly far apart on the same chromosome, and so are judged to be artifacts formed during construction of the library for sequencing. (C) Multihit. In this case, both reads in the pair have equally good alignments at multiple distinct locations.

on Golay codes with maximized edit distance, so that up to two errors in the index reads can be unambiguously corrected to recover the read.[56–58] Reads are subsequently trimmed to remove primer and LTR sequences (requiring exact matching to predicted sequences), yielding only genomic sequence data. A problem arises due to mispriming in the human genome, which can yield spurious integration sites. For this reason, we require a perfect match for the LTR segment extending between the 3′ end of the amplification primer to the 5′-CA-3′ sequence that defines the edge of the LTR.

Reads are next filtered to remove sequences complementary to the vector or virus used, requiring at least 75% global identity, a value chosen based on results with empirical datasets, and aligning in the first 5 nt of the read. Sequences are aligned to the reference genome using BLAT (parameters for alignment are found in the Materials and Methods).

Alignment information is then paired between the reads, and the integration site position and DNA fragment breakpoints (linker positions) are returned and stored in the IntSiteDB database (described in detail in our accompanying paper[52]). Read 1 and read 2 are joined based on location in the sequencing flow cell (encoded as the read name). Read pairs that map to identical sites are judged to be PCR duplicates and collapsed into single sites. To pass our quality filter, the genomic coordinates of these positions must lie within the range accessible by the sequencing chemistry—we allow a maximum of 2,500 base pairs (bp) as the default value (Figure 4A). Integration sites for which the read 1 (linker side) and read 2 (integration site side) positions are unreasonably distant, or on different chromosomes, are judged to be chimeras formed during PCR and are removed from the output data (Figure 4B). Paired alignments are additionally filtered for correct relative orientation.

Integration Sites in the Human Genome Showing Multiple Equally Good Alignments: "Multihits"

Viral or vector genomes that integrate within repetitive genomic elements often cannot be mapped to a single genomic coordinate, so that both read pairs align nearby, but they can be mapped to multiple locations in the human chromosomes (Figure 4C). For some forms of analysis, the multihits may be ignored—for example, in an analysis of integration site distributions relative to genomic features. However, for monitoring clinical gene therapy samples for possible adverse events, it is not safe to rule out possible insertional activation by integration in a repeated sequence—it is possible that an integration site in a multihit site may be near a cancer-related gene and involved in an adverse event. At least 40% of the human genome is composed of repeated sequences such as L1 retrotransposons, endogenous retroviruses, Alu elements, and others.[61]

A complication is that unique sonic fragments of the same parent integration site may map to non-identical lists of genomic coordinates, and even PCR duplicates may show different mapping behavior due to sequencing error. INSPIIRED thus uses a graph-theory-based approach to group alignments into clusters, so that each cluster can be treated as an integration site in downstream analysis. INSPIIRED assigns multihit reads to multihit clusters by creating an undirected graph $G = (V,E)$, where V is the set of reads identified as multihits and E is the set of pairs of multihit reads that share at least one putative integration site in the output list of multiple alignments. Each connected component of G is designated as a unique multihit cluster.

When considering the number of reads produced by the Illumina technology, the computational resources required to compare putative integration locations in a pairwise fashion can become prohibitive. To improve the scalability of multihit clustering, reads that have identical genomic DNA sequences across both read 1 and read 2 are combined into a single representative read before executing the pairwise comparison of potential genomic mappings.

Table 1. Processing of Unique In Silico-Generated Integration Sites Using INSPIIRED

	R2 (LTR Read) + R1 (Linker Read)				R2 (LTR Read) Only			
	0% Error	1% Error	2% Error	4% Error	0% Error	1% Error	2% Error	4% Error
Integration Sites								
Total simulated unique sites	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
Sites for which the collection of alignments contains the correct site	4,979	4,983	4,985	4,985	4,960	4,979	4,985	4,982
Site with single correct alignment	4,843	4,908	4,926	4,929	4,686	4,838	4,876	4,869
Sites with multiple alignments that include the correct site	136	189	144	96	276	334	264	197
Sites for which some read pairs show unique alignments while others show multiple alignments that include the correct alignment	617	613	532	347	477	547	483	291
Sites with no alignments	21	17	15	15	40	21	15	17
Sites for which individual reads yield different and/or incorrect alignment locations	87	229	278	179	75	222	280	198
Sequencing Reads								
Total simulated read pairs	500,000	500,000	500,000	500,000	500,000	500,000	500,000	500,000
Passed primer + LTRbit trimming	500,000	433,814	375,379	278,239	500,000	433,814	375,379	278,239
Passed linker trimming	500,000	433,797	375,059	275,280	500,000	433,797	375,059	275,277
Aligned correctly	489,927	406,195	313,977	113,580	486,939	404,667	313,387	113,617
Aligned unique integration site	458,457	384,179	299,833	109,808	450,702	379,411	296,961	109,155
Aligned multihit	31,470	22,016	14,144	3,772	36,237	25,256	16,426	4,462

When building an undirected graph from multihit read alignments, only the first connection of completely connected reads is used, reducing memory demand even further while yielding the same result with improved scalability.

Performance of the Pipeline Analyzed Using Synthetic Data

The performance of the pipeline was analyzed by generation and analysis of synthetic integration site data. Reads were generated with lengths of 179 and 143 nt corresponding to read 1 and read 2, respectively, including addition of the Illumina sequencing primers, DNA barcodes, primer landing pads, and flanking host DNA. A total of 5,000 sites were simulated. The distances between reads 1 and 2 were chosen randomly from a distribution of distances modeled to match empirical data, with 100 different distances between pairs sampled for each of the 5,000 integration sites. Four sets of the 5,000 integration sites were studied, containing no error, 1% error (roughly that expected from the Illumina sequencing method), 2% error, and 4% error.

Integration site datasets were trimmed, aligned, and quality filtered using the INSPIIRED pipeline. Results are tabulated in Table 1. Initially we asked whether each integration site could be recovered from at least one of the 100 read pairs (Table 1, top), and we then asked how many of the read pairs were recovered (Table 1, bottom).

For 0% error, 99.6% of sites could be recovered. Twenty-one sites were not aligned, and 87 sites were incorrectly aligned. These latter sites mapped to regions annotated as "low alignability," as defined by the GEM mappability program.[62] By visual inspection, these regions were rich in multiple repetitive element classes that were often nested within each other. Overall, of the 100 simulated sequence reads for each integration site, on average 98 could be mapped correctly. For the same integration sites containing 1%–4% error, the fractions mapping were 99.7% in each case.

The behavior of individual reads is summarized in the bottom of Table 1. Although the majority of sites were recovered, the proportion of the 100 paired sequences per integration site that could be recovered fell with increasing error, from 98% at 0% error to 23% at 4% error. Reassuringly, sequences lost with increasing amounts of error were mostly aligned correctly—error resulted in a lack of alignment, rather than misalignment.

We next asked how the sites were distributed among unique locations and repeated sequences. For each integration site in a repeated sequence, the R1 sequence at the linker end of each read has the potential of reaching into flanking unique DNA, resulting in a unique mapping position. At 0% sequence error, multihits accounted for only 6.4% of correctly aligned reads, while at 4% error, multihits accounted for 3.3%. Thus, the proportion of integration sites scoring as multihits was modest.

We investigated to what extent use of the linker side read (R1) allow for increased capture of unique sites and decreased recovery

Table 2. Experimental Datasets and Their SonicAbundance Values

Name	Description	Reads	Unique Sites	Multihits	Average SonicAbundance
LentiAcute	acute infection of HAP1 cells with a lentiviral vector	951,985	36,399	2,568	1.36
ClonedLenti[a]	mixture of cloned cells with five lentiviral integration sites	386,234	5	1	3,582.40
SCID, GTSP0855 p7 m162 PBMC	blood cell sample from the first SCID trial	845,267	666	89	40.17

[a]One of the ClonedLenti sites annotates as either a unique site or a multihit, depending on the length of the fragment recovered. Samples filtered on abundance of 20.

of multihits. We thus compared results with paired reads (R1 plus R2; Table 1, left side) to results with the integration site read only (R2; Table 1, right side). Results comparing R1 plus R2 to R2 alone showed a decrease in reads mapping to multiple locations (down 4,767 reads) and reads not aligning to the human genome (down 2,876 reads), as well as an increase in the number of reads that align correctly (2,988 more reads) and align to unique locations in the human genome (7,755 more reads). There was little difference in the number of sites detected, although there were half as many multihit sites when considering R1 plus R2 over R2 alone.

Analysis of Experimental Integration Site Data

We next assessed the performance of the pipeline over three empirical datasets (Table 2). The first consists of human HAP1 cells infected with an HIV-based vector and then grown for only 24 hr. A total of 38,967 integration sites were recovered, with an average of only 1.36 cells per site (SonicAbundance estimate). The second dataset consists of five purified 293T cell lines, each with a single lentiviral integration site. Cells were pooled and integration sites were determined from DNA purified from the mixture. Thus, only five sites were recovered by sequencing, with an average of 3,582 cells per site. The third dataset was a specimen of blood cells from a patient successfully treated for severe combined immunodeficiency (SCID-X1) using a gammaretroviral vector.[33] A total of 755 integration sites were recovered, ranging from 1 to 9,325 cells per site (average SonicAbundance of 40). This emphasizes that after gene correction, different progenitor cell clones delivered quite different proportions of cells to the periphery.

These data allow us to investigate the effects of human repeated sequences on integration site recovery using Illumina paired-end data. Multihits accounted for 6.6% of sites in the HAP1 cells and 11.8% in the SCID gene therapy specimen. Thus, multihits were roughly in the range expected from tests with the in silico-generated data.

Quantification was tested using the 293T cell clones with known integration sites mixed in different ratios. Figure 5 shows the expected and observed values (R = 0.852). In all cases, the expected ranking by frequency was observed, although there was some departure from the exact value expected. A clone included as only 1% of the population was readily detected.

DISCUSSION

Here, we present a pipeline for the generation and first-stage alignment of integration site data generated using Illumina paired-end reads. Portions of this pipeline have been used in earlier studies.[3,9,11,14,16–20,24,26,27,30,32,33,35,47,49–51,60,63–71] The full pipeline presented here introduces several improvements, including (1) use of a blocking oligonucleotide containing non-natural bases to suppress recovery of internal fragments; (2) use of all sets of paired-end reads, not just those that overlap, as in some earlier pipelines[43,45]; (3) a graph-theory-based method for tracking integration in repeated sequences; (4) use of synthetic data to assess the effect of read spacing and error; (5) implementation of the SonicAbundance method for quantification; and (6) use of cloned cell lines for empirical characterization of quantification accuracy.

Several integration site processing pipelines have been published, including IntegrationSeq/Map (2007), SeqMap (2008), QuickMap (2009), MAVRIC (2012), VISPA (2014), VISA (2015), and GeIST (2015).[29,36,38,39,42,43,46] Early pipelines focused on efficient mapping of single reads onto reference genomes[29,36,38] and included functions such as mapping the distance to the nearest transcription start site. Most pipelines include steps to reduce the total number of reads that must be aligned, so that computational time is used efficiently.

Illumina paired-end sequencing yields two reads from each DNA fragment, which may or may not overlap. When adapting paired-end sequencing to integration site pipelines, previous groups have chosen to merge these reads if possible,[39,42,43] but if they did not overlap, then the linker side read was discarded. INSPIIRED uses both reads, irrespective of overlap, for quantitative analysis of cell populations. That is, INSPIIRED uses all information to determine the location of integrated elements, and it also uses the location of the fragment breakpoint due to sonication for abundance quantification (Figure 3).

Determining the location of integration elements can be challenging due to repeated sequences in the human genome.[61] From our synthetic integration site data with 0% error, reads mapping to multiple locations in the human genome, "multihits," were found within repeated elements such as Alu elements, L1 retrotransposons, endogenous retroviruses, and others. Longer paired alignments to the human genome (that are generated by not requiring overlapping reads)

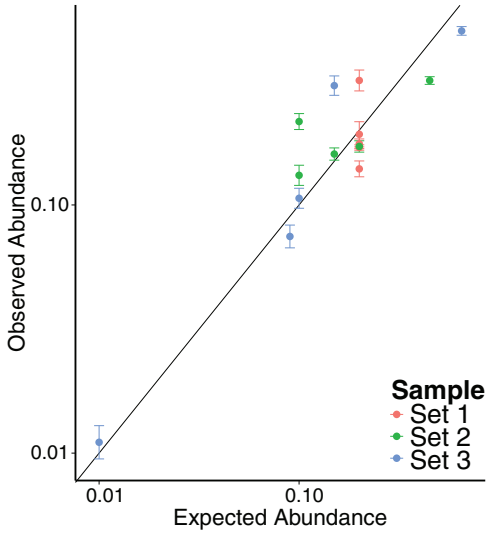

Figure 5. Results of a Control Study of Synthetic Mixtures of DNA from Cell Lines with Known Integration Site Distributions

The expected abundance based on the composition of the mixture is shown on the x axis, and the observed abundance (over 4 replicates) is shown on the y axis. Three cell mixtures were compared. In set 1, the clones were mixed in equal amounts. In set 2, the composition was 20% clone 1, 10% clone 2, 45% clone 3, 15% clone 4, and 10% clone 5. In set 3, the composition was 9% clone 1, 15% clone 2, 1% clone 3, 10% clone 4, and 65% clone 5. Error bars are SD.

have the potential to align the linker side read outside of these repeated regions, yielding a unique location for the integration site. This is evident in the analysis of in silico data (Table 1), in which more multihit sites were identified in paired reads (R1 plus R2) compared to those using the integration site sequence read only (R2). Multihit reads are grouped based on alignments and may have multiple linker attachment sites, as with unique integration sites. Thus, multihits can also be quantified by the SonicAbundance method and queried for possible clonal expansion.

The output data tables generated with INSPIIRED make possible the generation of a series of analytical products. These include (1) interactive heatmaps summarizing relationships of integration site data to genomic and epigenetic features; (2) reproducible reports on gene-corrected patient samples summarizing numerous features of integration site populations, including expansion of clones with integration sites near cancer-related genes; and (3) data frames suitable for statistical analysis based on the SonicAbundance method. These tools are described in our accompanying paper[52] and are available at https://github.com/BushmanLab/INSPIIRED.

MATERIALS AND METHODS

Recovery of Integration Sites Using PCR and a Blocking Primer

A detailed protocol is available in the Supplemental Materials and Methods, including the method for library preparation and sequencing of sites of new DNA integration in the human genome.

Samples of genomic DNA were prepared for Illumina sequencing by random shearing using a Covaris M220 ultrasonicator to achieve an average size distribution of 800–900 bp. DNA fragment ends were repaired (5′ phosphorylated and dA tailed) prior to TA ligation with custom linkers using NEBNext Ultra End Repair/dA-Tailing and NEBNext Ultra Ligation Modules, respectively (see Table S10 for linker design and Table S14 for sequences). Ligated DNA was split into at least four replicates prior to ligation-mediated (LM) PCR amplification (PCR1, 25 total cycles). Using the PCR1 product as a template, a nested LM PCR was conducted (PCR2, 20 total cycles) adding replicate-specific 12-bp Golay barcodes and Illumina adaptor sequences. Portions of PCR1 and PCR2 products were visually examined on ethidium bromide agarose gels. PCR2 products were pooled across replicates and bead purified prior to library construction. Sample concentrations were measured using the Quant-iT PicoGreen dsDNA Assay Kit and sequencing libraries were constructed by pooling samples by equal mass. Average amplicon size and library molarity were measured using an Agilent D1000 ScreenTape System and a KAPA SYBR FAST Universal qPCR Kit, respectively. Sequencing libraries were then sequenced on an Illumina MiSeq instrument. Sequences of oligonucleotides are provided in Tables S11 and S12.

Nested PCRs were supplemented with vector-specific blocking oligos complementary to the primer binding site found downstream of the 5′ LTR-U5 sequence. Blocking oligos contained nine blocked nucleic acids (BNAs) with a total length between 27 and 32 nucleotides and an estimated annealing temperature around 80°C. Each of the blocking oligos is terminated with a 3′ amino modification.

Paired-end sequencing was performed on the Illumina MiSeq, using 300-cycle V2 reagent kits with nano-, micro-, and standard flowcells. Cycle allocation was conducted as follows: read 1 used 179 cycles, index 1 used 12 cycles, and read 2 used 143 cycles. This allows for approximately 130 nucleotides of host sequence from both sides of the template molecule. Fastq output files were subsequently used as input for INSPIIRED.

Integration Site Quality Control and Read Trimming

Read sequences are trimmed to remove the linker, primer, and viral DNA end (LTRbit) sequences. A read pair is required to have the linker sequence at the beginning of read 1 and primer and LTRbit sequences at the beginning of read 2. The primer and the LTRbit sequences are determined by the corresponding vector sequence and are different for each virus or vector studied. Primer and LTRbit sequences are trimmed off from the beginning of read 2, and the linker sequences are trimmed off from read 1.

In cases where the sonic break point is close to the integration site, read 2 may read into the reverse complement of the linker and read 1 may read into the reverse complement of the primer and LTRbit. These sequences at the tails of the reads interfere with alignment and quality control, and thus are detected and trimmed off. About 20% of the reads are affected. These alignments are detected by

Bioconductor's Biostrings pairwiseAlignment function to trim and filter the leading and tailing sequences, while requiring 85% of identity based on edit distances.

Although a blocking oligo is employed to reduce pure vector sequence amplification during PCR, the blocking is not 100% efficient. Therefore, trimmed and filtered sequences are aligned to the vector reference and removed if either the read 1 or read 2 aligns with 75% of global identity and within 5 nt of the start of the read.

Sequence Alignment

INSPIIRED uses BLAT for DNA alignments because of its accuracy and the fact that it reports all alignments if a read has multiple hits in the genome with scores above a certain threshold, which makes it useful for handling multihit read pairs. The following BLAT parameters were used to align the read pairs: -tileSize = 11, -stepSize = 9, -minIdentity = 85, -maxIntron = 5, and -minScore = 27. Removing the commonly used -ooc = 11.ooc option led to better alignment in repeated regions, such as LINE, SINE, and LTR regions, since the option -ooc prohibits search in those regions. The option -maxIntron was changed to 5 from the default 750,000, as reads are amplified from genomic DNA and should not align across splicing elements. Tests showed that reducing the stepSize improved the accuracy of the alignments but increased the demand on memory; therefore, a stepSize of 9 represented a workable compromise. As BLAT is a local aligner, alignments were filtered out that only partially match to the host genome. Alignments were also filtered by a global identity score, defined as (matches + repMatches)/qSize for quality control on alignments, requiring a minimum of 95%. After aligning and filtering by the global identity score, the two reads of a template are paired by requiring that the reads (1) align to the same chromosome, (2) align to opposite strands, (3) maintain the correct predicted orientation (the linkered breakpoint read is "downstream" of the LTR read if the orientation is on the positive strand and vice versa for the negative strand), and (4) predicted template length is shorter than 2,500 nt.

For the studies reported here, the h18 draft of the human genome was used, to allow direct use of chromatin immunoprecipitation sequencing (ChIP-seq) data that was originally analyzed on this background (see our accompanying paper[52]). However, the organism and draft genome used for analysis can be selected by users and analyzed with any suitable annotation tracks compiled on that sequence.

The following definitions were used to describe the locus for each alignment in the format of "chromosome:strand:position." Here, tName, tStart, tEnd, and strand are columns of the BLAT output in the PSL format and they refer to chromosome, start position, end position, and strand, respectively. The integration site is given by [site = tName:+:tStart] if read 2 is mapped to the positive strand and [site = tName:-:tEnd] if mapped to the negative strand. Likewise, the break point is given by [breakpoint = tName:+:tEnd] if read 1 is mapped to the positive strand and [breakpoint = tName:-:tStart] if mapped to the negative strand. If more than one read pair yields the same integration site and breakpoint on the same strand, only one is kept and the others are considered to be PCR duplicates.

Generation of Synthetic Integration Site Data and Introduction of Error

Sequences were simulated from random locations in the human reference genome, then additional technical sequences required for the Illumina technology (Figure 1) were added to each read. Given a locus (composed of chromosome, strand, and location), the downstream sequence was obtained from the human reference genome. To generate simulated templates, the following components were concatenated together: Illumina P7, 12-nt Golay barcode, sequencing primer 2 (for both index and read 2 sequences), primer, LTRbit, reference genome sequence, linker (reverse complement), sequencing primer 1 (for read 1, reverse complement), and lastly, the Illumina P5 sequence (reverse complement). Read 2 sequences were generated by obtaining the 143 nt following the sequencing primer 2 of the above template. Read 1 sequences were generated by obtaining the 179 nt following the sequencing primer 1 after the reverse complementing the template. Should fewer bases exist in the template strand than are required, the ends of the reads are filled with poly(T) sequence, as is observed on the Illumina instrument. Index reads were generated as random 12 nt or a specified Golay DNA sequence. Sequencing primer, PCR primer, and LTRbit sequences are specific to each study.

For the synthetic dataset, 5,000 integration sites were generated from the human genome, and 100 lengths between read 1 and read 2 were simulated for each site. Simulated template lengths followed a normal distribution with a mean of 70 (SD of 250), while only keeping lengths greater than 30. In total, there are 500,000 read pairs in the simulation set. To evaluate the performance of INSPIIRED with sequencing error, four simulation datasets (with the same random sites and lengths) were generated containing 0%, 1%, 2%, and 4% read errors, applied after template construction.

Experimental Integration Site Analysis Workflow

After sequencing, Illumina output FASTQ files are used as inputs for INSPIIRED, along with sample information (which linkers and barcodes were used with which samples) and vector sequence information. As Golay DNA barcodes have large edit distances, the index sequence file that contains the barcode sequences for each read is subjected to error correction (up to two bases of correction). With the corrected barcodes, information from read 1 (containing the linker sequence) is used with the barcodes to demultiplex the samples, creating independent FASTQ files for read 1 and read 2. Sequences are then subjected to quality trimming (only keeping information with Q-scores greater than Q30). Following quality trimming, respective reads are filtered and trimmed for linker, primer, LTRbit, and vector-related sequences to yield only genomic DNA. The genomic sequences are then aligned to the host reference genome using BLAT, yielding an output PSL file with the alignment information. The alignment information is then used to determine the locations of alignments and read information is filtered and paired, as

previously discussed. Alignments for integration sites are then placed in four output files, including allSites (containing all uniquely mapped integration sites and their corresponding breakpoints), sites.final (contains a condensed form of allSites), multihitData (contains all properly paired and filtered integration sites that could not be uniquely mapped to the reference genome), and chimeraData (contains integration sites that failed proper pairing and are considered artifacts). These outputs are generated for each sample given a specific linker and barcode in the input sample information file.

Software Installation

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. The virtual machine, software, instructions, and a walkthrough that processes provided sample data are available at https://github.com/BushmanLab/INSPIIRED. As described in the instructions, the software supports two types of databases, MySQL and SQLite.

Data Availability

SRA accession numbers for the datasets studied can be found in Table S13.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Materials and Methods and fourteen tables and can be found with this article online at http://dx.doi.org/10.1016/j.omtm.2016.11.002.

AUTHOR CONTRIBUTIONS

E. Sherman, C.N., C.C.B., E. Six, M.C., and F.D.B. designed the study; E. Sherman, C.N., C.C.B., E. Six, M.C., F.M., S.R., M.J.D., P.B., S.H.-B.-A, L.C., and F.D.B. carried out the study; E. Sherman, C.N., C.C.B., E. Six, 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 from the NIH (AI052845, AI104400, AI082020, AI045008, AI117950, and HL113252), a grant from the European Research Council (ERC Regenerative Therapy 269037), and an award from the French National Agency for Research on AIDS and Viral Hepatitis. We also acknowledge support from the Penn Center for AIDS Research (P30-AI045008) and the PennCHOP Microbiome Program.

REFERENCES

1. Bushman, F.D. (2001). Lateral DNA Transfer: Mechanisms and Consequences (Cold Spring Harbor, NY: 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.W., 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. Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). Retroviruses (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).
5. 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.
6. 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.
7. 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.
8. Fischer, A., Hacein-Bey-Abina, S., and Cavazzana-Calvo, M. (2010). Gene therapy for primary immunodeficiencies. Immunol. Allergy Clin. North Am. *30*, 237–248.
9. 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.
10. Baum, C. (2007). Insertional mutagenesis in gene therapy and stem cell biology. Curr. Opin. Hematol. *14*, 337–342.
11. 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.
12. Kuo, C.Y., and Kohn, D.B. (2016). Gene therapy for the treatment of primary immune deficiencies. Curr. Allergy Asthma Rep. *16*, 39.
13. June, C.H., and Levine, B.L. (2015). T cell engineering as therapy for cancer and HIV: our synthetic future. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140374.
14. 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.
15. 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.
16. 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.
17. 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.
18. Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., et al. (2006). Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. *2*, e60.
19. Gijsbers, R., Ronen, K., Vets, S., Malani, N., De Rijck, J., McNeely, M., Bushman, F.D., and Debyser, Z. (2010). LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. Mol. Ther. *18*, 552–560.
20. 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.
21. 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.
22. Kustikova, O.S., Baum, C., and Fehse, B. (2008). Retroviral integration site analysis in hematopoietic stem cells. Methods Mol. Biol. *430*, 255–267.
23. Zychlinski, D., Schambach, A., Modlich, U., Maetzig, T., Meyer, J., Grassman, E., Mishra, A., and Baum, C. (2008). Physiological promoters reduce the genotoxic risk of integrating gene vectors. Mol. Ther. *16*, 718–725.

24. 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.

25. Kustikova, O.S., Schiedlmeier, B., Brugman, M.H., Stahlhut, M., Bartels, S., Li, Z., and Baum, C. (2009). Cell-intrinsic and vector-related properties cooperate to determine the incidence and consequences of insertional mutagenesis. Mol. Ther. *17*, 1537–1547.

26. 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.

27. 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.

28. 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.

29. Giordano, F.A., Hotz-Wagenblatt, A., Lauterborn, D., Appelt, J.U., Fellenberg, K., Nagy, K.Z., Zeller, W.J., Suhai, S., Fruehauf, S., and Laufs, S. (2007). New bioinformatic strategies to rapidly characterize retroviral integration sites of gene therapy vectors. Methods Inf. Med. *46*, 542–547.

30. Wang, G.P., Ciuffi, A., Leipzig, J., Berry, C.C., and Bushman, F.D. (2007). HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res. *17*, 1186–1194.

31. 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.

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. 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.

34. 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.

35. 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.

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

37. 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.

38. Appelt, J.U., Giordano, F.A., Ecker, M., Roeder, I., Grund, N., Hotz-Wagenblatt, A., Opelz, G., Zeller, W.J., Allgayer, H., Fruehauf, S., and Laufs, S. (2009). QuickMap: a public tool for large-scale gene therapy vector insertion site mapping and analysis. Gene Ther. *16*, 885–893.

39. Huston, M.W., Brugman, M.H., Horsman, S., Stubbs, A., van der Spek, P., and Wagemaker, G. (2012). Comprehensive investigation of parameter choice in viral integration site analysis and its effects on the gene annotations produced. Hum. Gene Ther. *23*, 1209–1219.

40. 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.

41. Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS ONE *8*, e64465.

42. 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.

43. 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.

44. Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. *7*, 2.

45. 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.

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

47. Meekings, K.N., Leipzig, J., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2008). HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog. *4*, e1000027.

48. Singh, P.K., Plumb, M.R., Ferris, A.L., Iben, J.R., Wu, X., Fadel, H.J., Luke, B.T., Esnault, C., Poeschla, E.M., Hughes, S.H., et al. (2015). LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. Genes Dev. *29*, 2287–2297.

49. 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.

50. 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.

51. 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.

52. Berry, C.C., Nobles, C., Six, E., Wu, Y., Malani, N., Sherman, E., Dryga, A., Everett, J.K., Male, F., Bailey, A., et al. (2017). INSPIIRED: quantification and visualization tools for analyzing integration site distributions. Moll Ther Methods Clin Dev. *4*, 17–26.

53. Braasch, D.A., and Corey, D.R. (2001). Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. Chem. Biol. *8*, 1–7.

54. Petersen, M., and Wengel, J. (2003). LNA: a versatile tool for therapeutics and genomics. Trends Biotechnol. *21*, 74–81.

55. Vester, B., and Wengel, J. (2004). LNA (locked nucleic acid): high-affinity targeting of complementary RNA and DNA. Biochemistry *43*, 13233–13241.

56. Hamady, M., Walker, J.J.J., Harris, J.K.J., Gold, N.J.N., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods *5*, 235–237.

57. 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.

58. 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.

59. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc. Natl. Acad. Sci. USA *108*, 20166–20171.

60. 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.

61. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome

Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

62. Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PLoS ONE *7*, e30377.

63. Lewinski, M.K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2005). Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. J. Virol. *79*, 6610–6619.

64. Ciuffi, A., Mitchell, R.S., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F.D. (2006). Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. Mol. Ther. *13*, 366–373.

65. Levine, B.L., Humeau, L.M., Boyer, J., MacGregor, R.R., Rebello, T., Lu, X., Binder, G.K., Slepushkin, V., Lemiale, F., Mascola, J.R., et al. (2006). Gene transfer in humans using a conditionally replicating lentiviral vector. Proc. Natl. Acad. Sci. USA *103*, 17372–17377.

66. 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.

67. 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.

68. Roth, S.L., Malani, N., and Bushman, F.D. (2011). Gammaretroviral integration into nucleosomal target DNA in vivo. J. Virol. *85*, 7393–7401.

69. Schaller, T., Ocwieja, K.E., Rasaiyaah, J., Price, A.J., Brady, T.L., Roth, S.L., Hué, S., Fletcher, A.J., Lee, K., KewalRamani, V.N., et al. (2011). HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. PLoS Pathog. *7*, e1002439.

70. Sharma, A., Larue, R.C., Plumb, M.R., Malani, N., Male, F., Slaughter, A., Kessl, J.J., Shkriabai, N., Coward, E., Aiyer, S.S., et al. (2013). BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc. Natl. Acad. Sci. USA *110*, 12036–12041.

71. Schneider, W.M., Brzezinski, J.D., Aiyer, S., Malani, N., Gyuricza, M., Bushman, F.D., and Roth, M.J. (2013). Viral DNA tethering domains complement replication-defective mutations in the p12 protein of MuLV Gag. Proc. Natl. Acad. Sci. USA *110*, 9487–9492.