

intmap: fast and flexible mapping of mobile DNA integration for basic and translational research

Gregory J. Bedwell^{1,2,*}, Peter Cherepanov³, Alan N. Engelman^{1,2}

¹Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA 02215, United States

²Department of Medicine, Harvard Medical School, Boston, MA 02115, United States

³Chromatin Structure and Mobile DNA Laboratory, The Francis Crick Institute, London, NW1 1AT, United Kingdom

*Corresponding author. Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA 02215, United States.

E-mail: gregoryj_bedwell@dfci.harvard.edu

Associate Editor: Jonathan Wren

Abstract

Summary: Integration of exogenous DNA into host cell chromatin is a hallmark of many virus and transposon lifecycles and the foundational basis for many modern cellular and genetic therapies. Defining integration sites (ISs) in a population of cells can accordingly inform fundamental aspects of mobile DNA biology and therapeutic treatment outcomes. Here, we describe intmap, a software that generalizes IS analysis to diverse data-types and experimental systems. intmap functions independently of strict library design assumptions and is highly tunable with respect to analysis parameters. Using both simulated and experimental data, we show that IS mapping with intmap is fast, accurate, and highly flexible.

Availability and implementation: intmap, related documentation, and a convenient installation script are available at <https://github.com/gbedwell/intmap>. Additional information is provided in the online Supplement. Analysis commands, outputs, and other information related to this manuscript are available at <https://doi.org/10.5281/zenodo.19475929>.

Introduction

Mobile DNA elements such as retroviruses and retrotransposons have significantly impacted Metazoan evolution (Craig *et al.* 2002). In these systems, integration can occur in numerous genomic locations, with multifactorial host-element interactions guiding the selection of particular target DNA acceptor sites (Kvaratskhelia *et al.* 2014, Sultana *et al.* 2017, Bourque *et al.* 2018, Jóźwik *et al.* 2022, Sabatino *et al.* 2022, Singh *et al.* 2022). Certain DNA viruses, such as adenoassociated virus (AAV), can additionally leverage double strand DNA breaks for integration (Miller *et al.* 2005, Hanlon *et al.* 2019, Breton *et al.* 2020, Nguyen *et al.* 2021, Van Gorder *et al.* 2023). The locations of integration events can profoundly impact cell biology, viral disease progression/persistence, cancer risk, therapeutic efficacy, and other medically-related outcomes (Howe *et al.* 2008, Johnson 2019, Nobles *et al.* 2020, Bedwell *et al.* 2021, Biasco *et al.* 2021, Coffin *et al.* 2021, Lian *et al.* 2021, Senft and Macfarlan 2021). As such, precise mapping of integration events is of interest in both basic and translational research.

Modern integration site (IS) mapping approaches utilize massively parallel sequencing platforms to simultaneously map thousands of events in a single experiment (Serrao *et al.* 2016, Afzal *et al.* 2017, Sherman *et al.* 2017, Spinozzi *et al.* 2017, Afzal

et al. 2020, Breton *et al.* 2020, Wells *et al.* 2020, Nguyen *et al.* 2021, Yan *et al.* 2023). IS libraries are most commonly prepared using ligation-mediated (LM) polymerase chain reaction (PCR), the closely related linear amplification-mediated (LAM)-PCR, or integration site loop amplification [ISLA; see Wells *et al.* (2020), Gabriel *et al.* (2014), and Wagner *et al.* (2014), respectively]. The final products of LM- and LAM-PCR are essentially identical: linear DNA molecules containing a terminal portion of the mobile element (i.e. the integrant), a portion of the host genome directly adjoining the integrant, and a specific linker sequence ligated to the opposing end of the genomic fragment (see Fig. 1, Step 1). Notably, the final library products for other sequencing approaches aimed at identifying chromosomal translocation hotspots (e.g., LAM-HTGTS) and sites of CRISPR/Cas off-target cleavages (e.g., GUIDE-seq, DISCOVER-seq, and CIRCLE-seq) share this same general structure (Tsai *et al.* 2015, 2017, Hu *et al.* 2016, Wienert *et al.* 2020). ISLA similarly produces linear DNA fragments that harbor the integrant terminus on one end. Unlike with LM- and LAM-PCR-based approaches, however, the non-integrant end harbors the reverse complement of a portion of the integrant terminus itself (Wagner *et al.* 2014). Other differences between IS library protocols can arise with respect to features such as the presence of unique molecular identifiers (UMIs), UMI structure and position, and the designed sequencing

Received: 5 December 2025. Revised: 15 April 2026. Accepted: 6 May 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

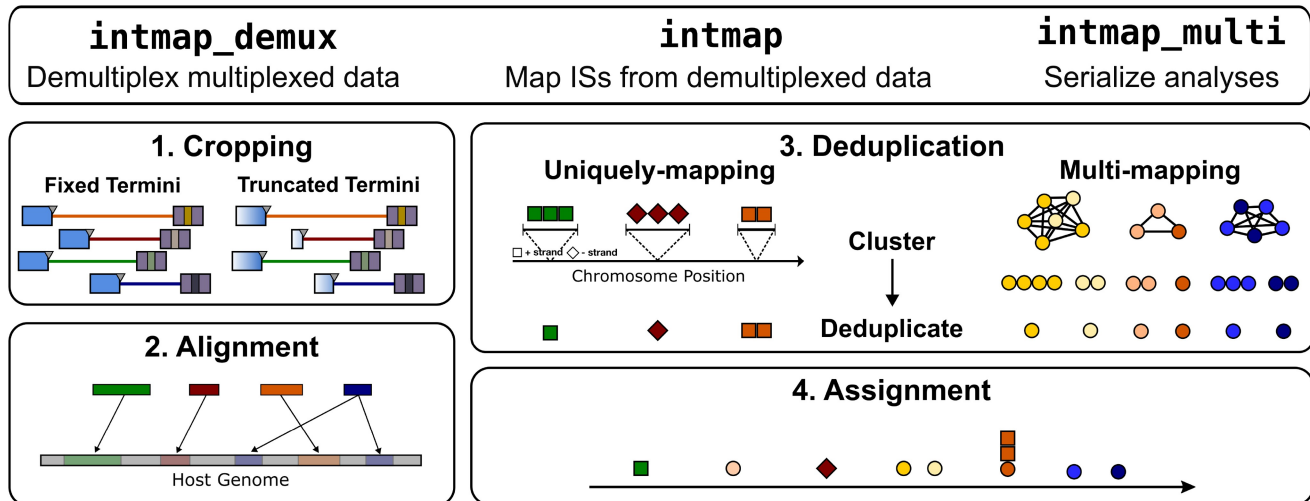


Figure 1 Schematic overview of IS mapping with intmap. The different intmap modules are listed in the top section. The four primary mapping steps—(i) cropping, (ii) alignment, (iii) deduplication, and (iv) assignment—are enumerated in the remaining sections. Cropping involves the removal of integrant and linker sequences (medium-blue and purple-gray boxes, respectively) from the sequenced fragments. Distinct UMIs are denoted as multi-colored bars within linkers. For integrant ends of variable length (referred to as truncated termini), the position along the integrant end is indicated according to the depth of color, with progression into darker colors indicating more full-length termini. Host genome fragments are represented as colored horizontal lines between integrant and linker sequences. The position of the integrant-host junctions (i.e. the genomic position that is ultimately mapped) are indicated by gray arrowheads. Alignment involves aligning cropped reads to the target genome. By default, intmap retains both uniquely-mapping (dark green, dark red, and dark orange bars) and multi-mapping (dark blue bar) fragments. Aligned fragments undergo quality control checks and are designated as either uniquely- or multi-mapping during these checks. For deduplication—the process of removing duplicates from the aligned fragments—similar fragments are clustered together according to either their aligned coordinates (uniquely-mapping fragments) or sequence similarity (multi-mapping fragments). Multi-mapping fragment clustering proceeds in multiple steps to sequentially reduce the search space for sequence similarity calculations. When UMIs are present, fragment clusters are sub-clustered based on UMI identity. A representative fragment from each (sub-)cluster is then chosen. In the deduplication diagram, uniquely-mapping fragments are represented as colored squares (plus strand fragments) or diamonds (minus strand fragments). Similar uniquely-mapping fragments are depicted in clusters originating from nearby genomic positions (brackets; uniquely-mapping fragment colors match Step 2 colors). Multi-mapping fragments are diagrammed as colored circles. For illustrative purposes, the multi-step clustering routine for multi-mapping fragments expands beyond the single multi-mapping fragment from Step 2 (dark blue bar/circles). Finally, retained fragments are assigned to their final genomic positions. Assignment criteria are further described in the text. Multi-mapping fragment reassignment is in part depicted as dark orange squares and circles being assigned to the same genomic position. Following mapping, auxiliary outputs, such as IS overlap with defined genomic features and peak-calling for Cas-mediated AAV integration hotspots are optionally available.

orientation of fragments [compare, e.g., the library designs described in [Serrao et al. \(2016\)](#), [Wells et al. \(2020\)](#), [Yan et al. \(2023\)](#), and [Sherman et al. \(2017\)](#)]. The utilization of different library architectures has led to the development of distinct analysis pipelines compatible with particular library designs but incompatible with others. These pipelines additionally differ in other key ways, such as the genome builds that they support, the data-types that they can analyze, the integrants that they can reliably map, and the manner in which they handle fragments that align equally well to multiple genomic positions (hereafter referred to as multi-mapping fragments).

Here, we describe intmap, a collection of Python-based command-line interface (CLI) modules intended to generalize IS analysis. intmap differs from existing IS mapping software in several important ways. First, intmap supports parallelization of resource-intensive tasks across Unix-based systems. Second, intmap is highly flexible with respect to the data-types and library designs that it accepts. Third, intmap can accurately map IS locations of diverse integrant species, including retroviruses and AAV, which differ markedly in the nature of their terminal sequences. Fourth, intmap enables the tunability of a large number of mapping parameters to

meet diverse analytical needs. Fifth, intmap efficiently and scalably processes and retains both uniquely- and multi-mapping fragments and assigns multi-mapping fragments to probable genomic locations without requiring manual inspection. Finally, intmap incorporates both new and previously described methods to minimize experimental artifacts.

Implementation

intmap's functionality is implemented across three separate CLI modules: intmap_demux for demultiplexing multiplexed data, intmap for mapping ISs to genomic positions, and intmap_multi for automated serialization of multiple analyses ([Fig. 1](#)). External software requirements are listed in [Section 1.1 of the Supplementary Material](#).

Demultiplexing

intmap_demux can demultiplex multiplexed data based on sample-specific in-line indexes (barcodes), header-based

indexes, or separate index files. See [Section 1.2 of the Supplementary Material](#) for additional information.

Mapping

Mapping of demultiplexed reads to genomic locations proceeds in four key steps: cropping, alignment, deduplication, and assignment ([Fig. 1](#)). Each of these are individually described in the sections below.

Cropping

Techniques used for IS library generation rely on known sequences on both termini of DNA fragments for PCR amplification. These sequences must be identified and removed from each read in order to accurately define integrant-host junctions ([Fig. 1](#), Step 1). *intmap*'s cropping routine identifies user-defined end sequences (e.g., integrant and linker sequences) in each read in the input file(s). When these sequences are found, the sequence of the given read up to and including the match positions are removed and the remaining sequence is saved to a new FASTQ file. UMIs are extracted during cropping. When UMIs are absent, they are defined as “N” for use in downstream processing. Additional information is given in [Section 1.3 of the Supplementary Material](#).

Alignment

Alignment proceeds in three stages. Reads are first aligned to a user-defined integrating-element sequence (retrovirus, transposon, etc.) to identify and remove reads arising from internal priming, plasmid contamination, or auto-integration. Reads that align in this step are discarded. Reads that pass step 1 are aligned to a user-defined reference genome ([Fig. 1](#), Step 2). For short-reads, Bowtie2's end-to-end algorithm is used for alignment ([Langmead and Salzberg 2012](#)), while long-read alignment utilizes minimap2 ([Li 2018](#)). Finally, aligned reads are processed for quality. The user is able to define the minimum- and maximum-allowed fragment lengths, the maximum mismatch rate, the maximum indel rate, the minimum mapping quality (MAPQ) score, the minimum average Phred quality score, and the maximum number of bases on the integrant-end before which there is an exact match to the reference genome. In *intmap*'s quality control (QC) routine, soft-clipped and ambiguous bases are treated as mismatches. Only aligned reads that pass the QC checks are carried on for further processing. Aligned fragments are stratified as uniquely- or multi-mapping during QC ([Fig. 1](#)).

Deduplication

Deduplication procedures differ for uniquely- and multi-mapping fragments. For the former, fragments with identical alignment positions and UMIs are first grouped together and the fragment with the highest MAPQ and average Phred scores is chosen as the representative fragment. Next, “fuzzy” deduplication groups together uniquely-mapping fragments whose start and end coordinates fall within a user-defined window ([Fig. 1](#), Step 3). A similar approach is

used in [Wells et al. \(2020\)](#). The UMIs of grouped fragments are then compared using a directional clustering algorithm similar to the “directional” algorithm in UMI-tools ([Smith et al. 2017](#)). More information on UMI deduplication is given in [Section 1.4.1 of the Supplementary Material](#). For multi-mapping fragments, the aligned fragment position is uninformative; the sequences themselves are the only reliable source of information. Multi-mapping fragment deduplication begins by collapsing fragments with the same sequences and UMIs into a single representative sequence. For paired-end reads, concatenated read-pairs are used for sequence comparisons. Next, representative fragments are compared for sequence similarity ([Fig. 1](#), Step 3). *intmap* attempts to minimize the search space for each sequence by clustering similar sequences together prior to quantifying exact sequence similarity ([Fig. 1](#), Step 3; see [Section 1.4.2 of the Supplementary Material](#) for more information). Ultimately, sufficiently similar fragments are grouped together as fuzzy matches, and UMI deduplication proceeds as for uniquely-mapping fragments. Only deduplicated fragments are carried on for further processing ([Fig. 1](#), Step 3). These deduplication procedures work similarly on data with and without UMIs, providing a consistent deduplication workflow for IS analysis independent of library design.

Assignment

Following deduplication, fragments are assigned to their final genomic locations ([Fig. 1](#), Step 4). Uniquely-mapping fragments are assigned to their aligner-designated positions. For multi-mapping fragments, *intmap* can assign final positions in two distinct ways. The first way, which works similarly to the approach for uniquely-mapping fragments, is intended for situations where multiple copies of the same insertion (e.g., in clonally expanded cell populations or site-specific integration) are unexpected. The aligners used in *intmap* randomly choose one of the possible best-scoring alignment positions as the primary multi-mapping fragment position. While the ISs derived from these randomly chosen positions are not unambiguously defined, their retention works to accurately reflect the degree of integration stemming from multi-mapping fragments in annotated genomic regions (see [Section 2.1.1 of the Supplementary Material](#)). With 53.9% of the human genome classified as repetitive in the telomere-to-telomere assembly, the ability to quantify integration within repetitive regions is critical to obtain comprehensive views of integration across different contexts/systems ([Hoyt et al. 2022](#)). This approach to multi-mapping fragment position assignment, however, may under-represent any clonal or site-specific IS populations that are partially or completely composed of multi-mapping fragments. To address this limitation, *intmap* can alternatively assign multi-mapping fragment positions based on sequence similarity with uniquely-mapping fragments or other multi-mapping fragments with distinct UMIs and/or linker ligation positions (“multi-mapping fragment reassignment”; see [Section 1.5.1 of the Supplementary Material](#) for more information). This assignment mechanism, suitable when clonal expansion or site-specific integration are expected, helps to accurately represent positionally-biased IS populations in the output (see [Section 2.1.2 of the Supplementary Material](#)). When all fragments have been processed, *intmap* makes a final

pass through the assigned fragment positions to tidy the final data, which is described in more detail in [Section 1.5.2 of the Supplementary Material](#).

Multiple analyses

intmap_multi serializes multiple analyses in a single command. It is written to be highly modular, allowing different analysis parameters to be defined within the same set of serialized analyses. intmap's GitHub repository contains more information on the structure of the input file used by intmap_multi. Additional examples are provided on Zenodo.

Results

intmap accurately mapped ISs across diverse data-types, library designs, and integrant-types ([Figures S1C-E](#), S2B-D, S3B-C, S4A-B, and S5A-C). These data-types included paired-end short-read, single-end short-read, and long-read data ([Figures S1C-E](#), S2B-D, S3B-C, S4A-B, and S5A-C). The integrant-types included fixed-end integrants (e.g., retroviruses; [Figures S1C-E](#), S2B-D, S3B-C, and S4A-B) and randomly truncated integrants (e.g., AAV; [Figures S5A-C](#)). Parallelization of analyses across multiple cores decreased runtime several-fold ([Figure S1A](#)). intmap was more-over ~6-fold faster than the state-of-the-art IS mapping software INSPIRED, while maintaining mapping accuracy ([Figure S3A-C](#)). Finally, intmap efficiently and robustly handled multi-mapping fragments, mapping multi-mapping fragments without substantially impacting runtime and correctly identifying clonal populations comprised entirely of multi-mapping fragments ([Figures S1A-D and S2A-D](#)).

Summary

Altogether, intmap generalizes IS analyses. intmap's agnostic approach to library design specifications works to democratize IS analysis, freeing users from rigidly defined read structures and providing a consistent framework for robust IS analysis across systems. It moreover enhances the possibility of cross-study comparisons and meta-analyses of IS data. The tunability of intmap's analysis parameters additionally provides analytical flexibility to better meet varying analytical needs. Finally, intmap's applicability to diverse integrant-types establishes it as a proverbial Swiss Army knife for IS analysis. We expect that intmap will become a valuable tool for research across fields related to genomic editing and genomic integration.

Acknowledgement

We thank Dr. Parmit K. Singh for independently testing intmap.

Author contributions

Gregory Bedwell (Conceptualization [lead], Formal analysis [lead], Methodology [lead], Software [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]),

Peter Cherepanov (Methodology [supporting], Software [supporting], Visualization [supporting], Writing—review & editing [supporting]), and Alan Engelman (Supervision [lead], Visualization [supporting], Writing—review & editing [equal])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by grants from the National Institutes of Health to A.N.E. [R37AI039394, R01AI052014]. P.C. is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK [CC2058], the UK Medical Research Council [CC2058], and the Wellcome Trust [CC2058].

Data availability

intmap is available at <https://github.com/gbedwell/intmap>. Small example datasets, along with analysis commands, and enumeration of all accepted arguments, are provided on GitHub. Commands and output for other analyses described here, along with related information, are available at <https://doi.org/10.5281/zenodo.19475929>.

References

- Afzal S, Wilkening S, Von Kalle C *et al*. GENE-IS: time-Efficient and accurate analysis of viral integration events in Large-Scale gene therapy data. *Mol Ther Nucleic Acids* 2017;**6**:133–9. <https://doi.org/10.1016/j.omtn.2016.12.001>
- Afzal S, Fronza R, Schmidt M. VSeq-Toolkit: comprehensive computational analysis of viral vectors in gene therapy. *Mol Ther Methods Clin Dev* 2020;**17**:752–7. <https://doi.org/10.1016/j.omtm.2020.03.024>
- Bedwell GJ, Jang S, Li W *et al*. Rigrag: high-resolution mapping of genic targeting preferences during HIV-1 integration *in vitro* and *in vivo*. *Nucleic Acids Res* 2021;**49**:7330–46. <https://doi.org/10.1093/nar/gkab514>
- Biasco L, Izotova N, Rivat C *et al*. Clonal expansion of T memory stem cells determines early anti-leukemic responses and long-term CAR T cell persistence in patients. *Nat Cancer* 2021; **2**:629–42. <https://doi.org/10.1038/s43018-021-00207-7>
- Bourque G, Burns KH, Gehring M *et al*. Ten things you should know about transposable elements. *Genome Biol* 2018;**19**:199. <https://doi.org/10.1186/s13059-018-1577-z>
- Breton C, Clark PM, Wang L *et al*. ITR-Seq, a next-generation sequencing assay, identifies genome-wide DNA editing sites *in vivo* following adeno-associated viral vector-mediated genome editing. *BMC Genomics* 2020;**21**:239. <https://doi.org/10.1186/s12864-020-6655-4>

- Coffin JM, Bale MJ, Wells D *et al.* Integration in oncogenes plays only a minor role in determining the *in vivo* distribution of HIV integration sites before or during suppressive antiretroviral therapy. *PLoS Pathog* 2021;**17**:e1009141. <https://doi.org/10.1371/journal.ppat.1009141>
- Craig NL, Craigie R, Gellert M, editors *et al.* *Mobile DNA II*. Washington, D.C.: ASM Press, 2002.
- Gabriel R, Kutschera I, Bartholomae CC *et al.* Linear amplification mediated PCR – localization of genetic elements and characterization of unknown flanking DNA. *J Vis Exp* 2014;**88**:e51543. <https://doi.org/10.3791/51543>
- Hanlon KS, Kleinstiver BP, Garcia SP *et al.* High levels of AAV vector integration into CRISPR-induced DNA breaks. *Nat Commun* 2019;**10**:4439. <https://doi.org/10.1038/s41467-019-12449-2>
- Howe SJ, Mansour MR, Schwarzwaelder K *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J Clin Invest* 2008;**118**:3143–50. <https://doi.org/10.1172/JCI35798>
- Hoyt SJ, Storer JM, Hartley GA *et al.* From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* 2022;**376**:eabk3112. <https://doi.org/10.1126/science.abk3112>
- Hu J, Meyers RM, Dong J *et al.* Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat Protoc* 2016;**11**:853–71. <https://doi.org/10.1038/nprot.2016.043>
- Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* 2019;**17**:355–70. <https://doi.org/10.1038/s41579-019-0189-2>
- Jóźwik IK, Li W, Zhang D-W *et al.* B-to-a transition in target DNA during retroviral integration. *Nucleic Acids Res* 2022;**50**:8898–918. <https://doi.org/10.1093/nar/gkac644>
- Kvaratskhelia M, Sharma A, Larue RC *et al.* Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res* 2014;**42**:10209–25. <https://doi.org/10.1093/nar/gku769>
- Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9. <https://doi.org/10.1038/nmeth.1923>
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. <https://doi.org/10.1093/bioinformatics/bty191>
- Lian X, Gao C, Sun X *et al.* Signatures of immune selection in intact and defective proviruses distinguish HIV-1 elite controllers. *Sci Transl Med* 2021;**13**:eabl4097. <https://doi.org/10.1126/scitranslmed.abl4097>
- Miller DG, Trobridge GD, Petek LM *et al.* Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. *J Virol* 2005;**79**:11434–42. <https://doi.org/10.1128/JVI.79.17.11434-11442.2005>
- Nguyen GN, Everett JK, Kafle S *et al.* A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. *Nat Biotechnol* 2021;**39**:47–55. <https://doi.org/10.1038/s41587-020-0741-7>
- Nobles CL, Sherrill-Mix S, Everett JK *et al.* CD19-targeting CAR T cell immunotherapy outcomes correlate with genomic modification by vector integration. *J Clin Invest* 2020;**130**:673–85. <https://doi.org/10.1172/JCI130144>
- Sabatino DE, Bushman FD, Chandler RJ, American Society of Gene and Cell Therapy (ASGCT) Working Group on AAV Integration *et al.* Evaluating the state of the science for adeno-associated virus integration: an integrated perspective. *Mol Ther* 2022;**30**:2646–63. <https://doi.org/10.1016/j.ymthe.2022.06.004>
- Senft AD, Macfarlan TS. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* 2021;**22**:691–711. <https://doi.org/10.1038/s41576-021-00385-1>
- Serrao E, Cherepanov P, Engelman AN. Amplification, next-generation sequencing, and genomic DNA mapping of retroviral integration sites. *J Vis Exp* 2016;**109**:e53840. <https://doi.org/10.3791/53840>
- Sherman E, Nobles C, Berry CC *et al.* INSPIRED: a pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. *Mol Ther Methods Clin Dev* 2017;**4**:39–49. <https://doi.org/10.1016/j.omtm.2016.11.002>
- Singh PK, Bedwell GJ, Engelman AN. Spatial and genomic correlates of HIV-1 integration site targeting. *Cells* 2022;**11**:655. <https://doi.org/10.3390/cells11040655>
- Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;**27**:491–9. <https://doi.org/10.1101/gr.209601.116>
- Spinozzi G, Calabria A, Brasca S *et al.* VISPA2: a scalable pipeline for high-throughput identification and annotation of vector integration sites. *BMC Bioinformatics* 2017;**18**:520. <https://doi.org/10.1186/s12859-017-1937-9>
- Sultana T, Zamborlini A, Cristofari G *et al.* Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* 2017;**18**:292–308. <https://doi.org/10.1038/nrg.2017.7>
- Tsai SQ, Zheng Z, Nguyen NT *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* 2015;**33**:187–97. <https://doi.org/10.1038/nbt.3117>
- Tsai SQ, Nguyen NT, Malagon-Lopez J *et al.* CIRCLE-seq: a highly sensitive *in vitro* screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat Methods* 2017;**14**:607–14. <https://doi.org/10.1038/nmeth.4278>
- Van Gorder L, Doshi BS, Willis E *et al.* Analysis of vector genome integrations in multicentric lymphoma after AAV gene therapy in a severe hemophilia A dog. *Mol Ther Methods Clin Dev* 2023;**31**:101159. <https://doi.org/10.1016/j.omtm.2023.101159>
- Wagner TA, McLaughlin S, Garg K *et al.* Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* 2014;**345**:570–3. <https://doi.org/10.1126/science.1256304>
- Wells DW, Guo S, Shao W *et al.* An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. *BMC Genomics* 2020;**21**:216. <https://doi.org/10.1186/s12864-020-6647-4>
- Wienert B, Wyman SK, Yeh CD *et al.* CRISPR off-target detection with DISCOVER-seq. *Nat Protoc* 2020;**15**:1775–99. <https://doi.org/10.1038/s41596-020-0309-5>
- Yan A, Baricordi C, Nguyen Q *et al.* Is-Seq: a bioinformatics pipeline for integration sites analysis with comprehensive abundance quantification methods. *BMC Bioinformatics* 2023;**24**:286. <https://doi.org/10.1186/s12859-023-05390-1>

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Bioinformatics, 2026, 42, 1–5

<https://doi.org/10.1093/bioinformatics/btag310>

Applications Note