Shao et al. Retrovirology (2016) 13:47
DOI 10.1186/s12977-016-0277-6

Retrovirology

SHORT REPORT Open Access

Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes

Wei Shao[1*], Jigui Shan[1], Mary F. Kearney[2], Xiaolin Wu[3], Frank Maldarelli[2], John W. Mellors[4], Brian Luke[1], John M. Coffin[5] and Stephen H. Hughes[2]

Abstract

The NCI Retrovirus Integration Database is a MySql-based relational database created for storing and retrieving comprehensive information about retroviral integration sites, primarily, but not exclusively, HIV-1. The database is accessible to the public for submission or extraction of data originating from experiments aimed at collecting information related to retroviral integration sites including: the site of integration into the host genome, the virus family and subtype, the origin of the sample, gene exons/introns associated with integration, and proviral orientation. Information about the references from which the data were collected is also stored in the database. Tools are built into the website that can be used to map the integration sites to UCSC genome browser, to plot the integration site patterns on a chromosome, and to display provirus LTRs in their inserted genome sequence. The website is robust, user friendly, and allows users to query the database and analyze the data dynamically. Availability: https://rid.ncifcrf.gov; or http://home.ncifcrf.gov/hivdrp/resources.htm.

Keywords: Retrovirus, HIV, Integration site, Database, Integration site assay, ISA, Expanded clones

Background

For a retrovirus to replicate, the virus must integrate a DNA copy of its genome, producing a provirus in the genome of the infected host cell. Research into host integration sites of retroviral genomes has been on-going for many years [2, 8, 13, 14]. Insertion into regions near host genes can affect the expression of the host gene. If the host gene has an important role in controlling cell growth and division, integration can cause clonal cell expansion, and may be involved in the development of malignancy [1, 12, 15, 18].

The advent of next generation sequencing technologies has allowed for tens of thousands or even millions of retroviral integration sites to be obtained in single experiments [5, 10, 16, 19, 20]. Currently, integration site information must be downloaded from the supplementary files of publications or obtained from the investigators directly, making collection time consuming and difficult. Recently, there has been a rapid increase in the amount of retroviral integration site data that is available, and there is a need for a readily accessible database to store, retrieve, and analyze integration site data. In addition, a public integration site database will allow concurrent mapping and reporting of proviral orientation across and among studies, and can help to avoid issues that can arise when integration sites are mapped using different genome builds or by applying different definitions for the orientation of the gene or the provirus. For example, Maldarelli et al. and Wagner et al. mapped integration sites to the human genome build hg19 [12, 18], whereas Ikeda et al. and Wang et al. mapped their integration sites to an older genome build [9, 19]. Furthermore, LaFave et al. and Wagner et al. defined "+" proviral orientation as being in the same orientation as the chromosome [10, 18], whereas Han et al. [7] and Sunshine et al. [17] defined "+" proviral orientation as being the same as the target

*Correspondence: shaow@mail.nih.gov
[1] Advanced Biomedical Computing Center, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research (FNLCR), Frederick, MD, USA
Full list of author information is available at the end of the article

gene. To avoid such inconsistencies and to facilitate the storage, retrieval, and coordinated analyses of published retroviral integration site data, we built the NCI Retrovirus Integration Database (RID) (https://rid.ncifcrf.gov/, Fig. 1) and are making it available for public use.

Methods

We collected retrovirus integration sites information from published papers or by directly contacting the authors when the information that was not readily available in the published papers (see acknowledgements). For consistency, we only extracted host, chromosome, integration site, virus type or subtype, proviral orientation, and LTR from those datasets and then we performed gene mapping (including intron/exon mapping) using NCBI genome. This local gene annotation database is derived from NCBI genomes (http://www.ncbi.nlm.nih.gov/genome/). If an integration site is not in a gene, then the nearest genes in both directions were mapped and stored in RID. All gene annotations were based on human genome build GRCH37/hg19. For the raw data using older genome builds, the integration sites were converted to hg19 using LiftOver, a genome converting tool provided by UCSC Genome Bioinformatics (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Proviruses orientations have been converted to a uniform standard: if a provirus is integrated in the same orientation as the target chromosome (using the UCSC numbering convention), it is defined as "+", otherwise, it is defined as "−".

Results

RID provides a common place to store and retrieve information describing retroviral integration sites. It is intended for public use and requires no login information. The database stores information on the sites of retroviral integrations into host genomes, the host type, virus type and subtype, a description of the sample origin, such as tissue type, and the reference from which the data originated. The integration site information is presented in a table that includes the host chromosome number, the specific coordinates of integration, the nearest gene, whether the integration site was identified from the retroviral 5′LTR or 3′LTR; and, if the integration site is in a gene, whether it is in an exon or an intron. Currently, RID includes valid data from retroviral insertion sites of HIV-1, HTLV-1, and MLV from multiple publications [4, 5, 7, 9–12, 14, 16, 18] and the database is intended to include integration site information from other retrovirus as more data become available. All of the data in RID have been mapped to a recent completely annotated genome build for the specific host, for example, human genome hg19 for HIV-1 and HTLV-1.

Accessing information on the database

The database can be accessed using current version of web browsers including Internet Explorer, Chrome, Firefox, and Safari. It is compatible with PC, Mac, iPad, and cellphones. The main menu for the RID web interface is divided into five sections (Fig. 1): Choose virus and subtype, Choose host and chromosomes, Query options, Integration site information selection, and Advanced queries. The main menu allows users to access data by searching for integration sites for a specific virus or a specific viral subtype in the "Choose virus and subtype" section. Users then can access the data by selecting a specific host type and one or all of the chromosomes from "Choose host and chromosomes" section. Users can then select the "Submit Query" button to display the query result.

Users can limit their query by choosing an option in the "Query option" section. For example, a nucleotide position range on a specific chromosome can be chosen to search for integration sites within a specific region of the host genome or users can search query integration sites based on genes, the PubMed ID of one or two specific publications, or a sample name or a tissue type to narrow the query. The "ADVANCED QUERIES" section can be used to find integrations that have been reported in the same genes across multiple studies.

The results of any search can be exported, as a text file (Fig. 2), for inclusion in presentations or publications. After obtaining query results, users can click the "I" button on the results page (Fig. 2) to display the chromosome information for the integration sites including the sequence data for the 500 host nucleotides flanking the integration site (Fig. 3) joined to a fragment of nucleotides at each end of the consensus LTR for the virus chosen. It also shows the correct length of the target site duplication depending on the virus; for example, for HIV-1, it shows five nucleotide duplications, for HTLV-1, it shows six nucleotide duplications, and for MLV, it shows four nucleotide duplications at each end the provirus [2]. In this display, the 5′LTR is highlighted in red and the 3′LTR in blue. Users can also click the "G" button on the results page (Fig. 2) to display a particular integration site relative to the full chromosome on the UCSC genome page (https://genome.ucsc.edu/, Fig. 4a) or they can click the hyperlink to "gene_id" to display the detailed gene information from the NCBI Gene database (http://www.ncbi.nlm.nih.gov/gene/). The "pubmed_id" link will provide the corresponding paper from NCBI PubMed.

RID also includes tools to show the distribution of integration sites along a chromosome. After choosing a chromosome, users can click the "Genome mapping" button to display all the integration sites mapped to a specific chromosome in UCSC genome browser. They can also

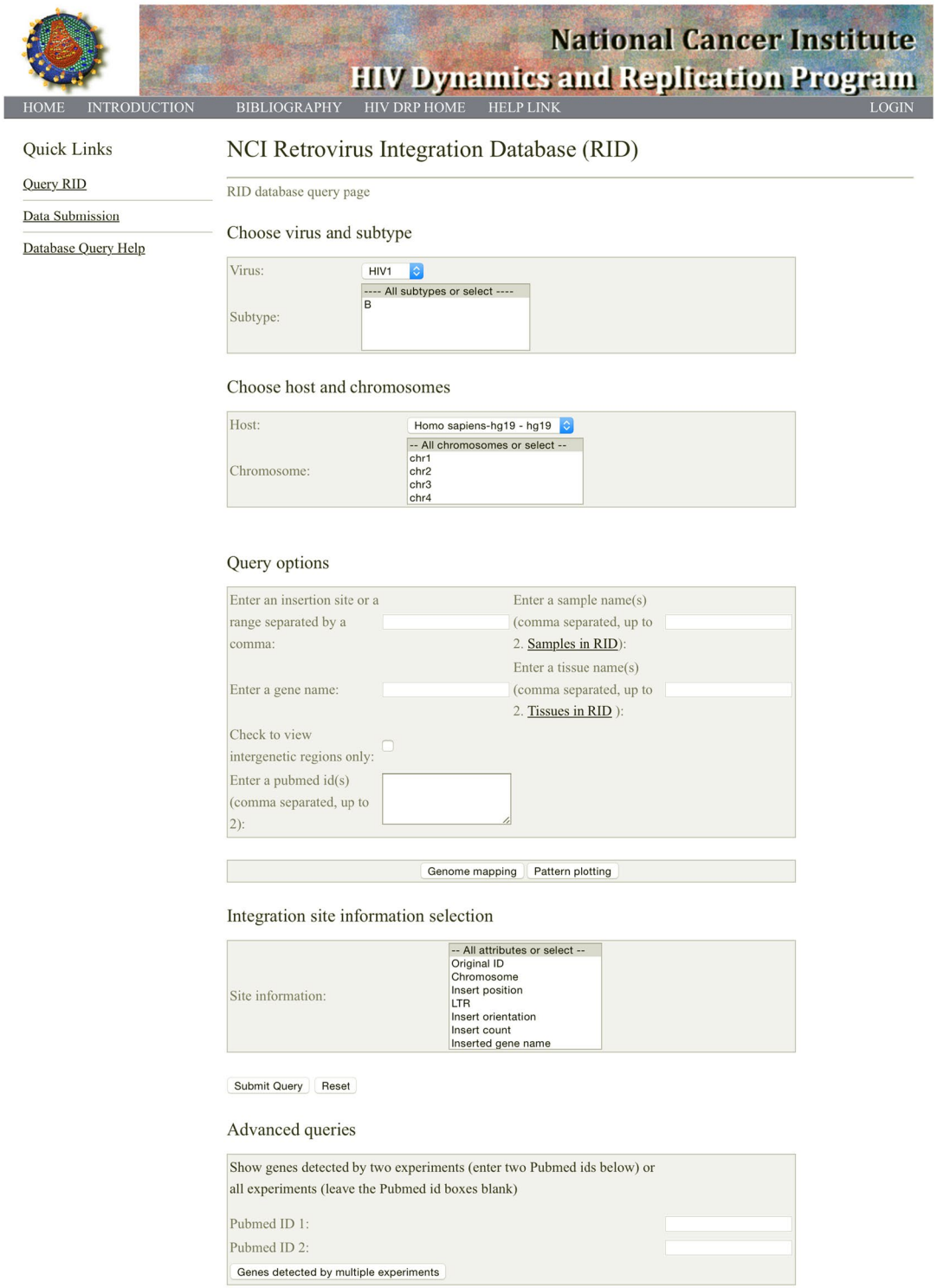

Fig. 1 Screen shot of the RID home page

RID Query Result

I - Show provirus orientation and integration
G - Map to ucsc genome browser

Download Result

origin_id	chr	insert_position	LTR	insert_orientation	inserted_gene	refseq_id	gene_id	gene_orientation	exon_intron	nearest_gene	nearest_gene_refseq_id	nearest_gene_id	nearest_gene_orientation	nearest_gene_distance	comment	pubmed_id
21	chr1	8448583 I G	5LTR	+	RERE	NM_001042681	473	-	intron12							15163705
21	chr1	10034649 I G	5LTR	+	NMNAT1	NM_022787	64802	+	intron2							15163705
149	chr1	23891195 I G	5LTR		not found					ID3	NM_002167	3399	-	4910		15163705
22	chr1	35950701 I G	5LTR	-	KIAA0319L	NM_024874	79932	-	intron3							15163705
21	chr1	35978232 I G	5LTR	+	KIAA0319L	NM_024874	79932	-	intron2							15163705
150	chr1	64987369 I G	5LTR	-	CACHD1	NM_020925	57685	+	intron1;Flag=First							15163705
144	chr1	153866184 I G	5LTR	+	GATAD2B	NM_020699	57459	-	intron1;Flag=First							15163705

Fig. 2 Partial screen shot of query results from the RID. In each *row*, clicking buttons "I", "G", or hyperlinks for gene_id, and pubmed_id can be used to link the integration site being investigated to the corresponding host genome sequence, host genome mapping, gene information, and PubMed abstract

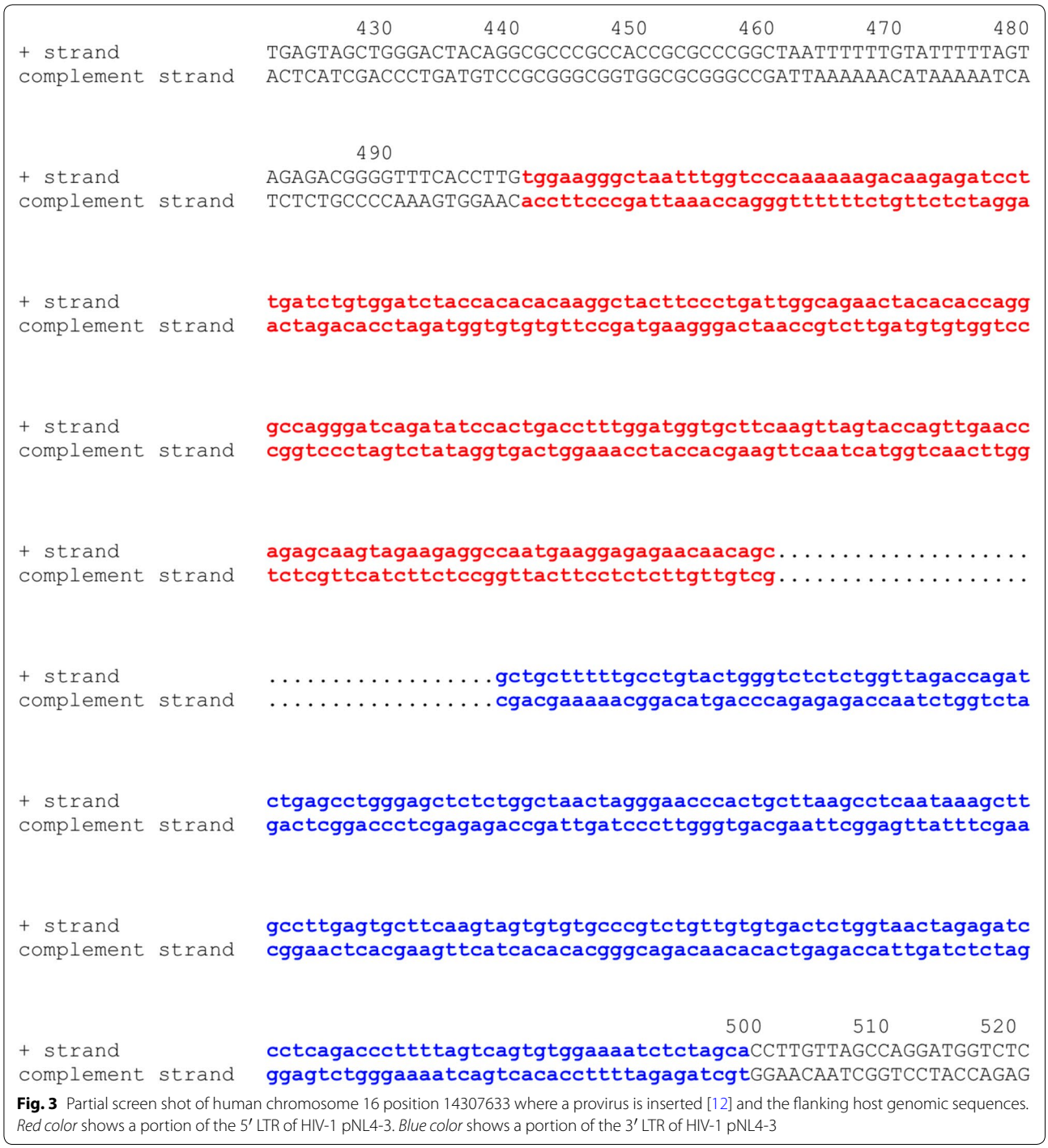

Fig. 3 Partial screen shot of human chromosome 16 position 14307633 where a provirus is inserted [12] and the flanking host genomic sequences. *Red color* shows a portion of the 5′ LTR of HIV-1 pNL4-3. *Blue color* shows a portion of the 3′ LTR of HIV-1 pNL4-3

combine three query options; for example, chromosome 17, gene name STAT5B, and Pubmed_id 24968937 [12], to map the integration sites in STAT5B in UCSC genome browser (Fig. 4b) or they can click "Pattern plotting" to display the distribution of the integration sites on specific chromosomes in 1 million nucleotide bins (Fig. 5a) by, for example, selecting chromosome 22. Note that no integration sites in chromosome 22 are seen in the first 14 million bases or so, reflecting the fact that chromosome 22 is one of the five human acrocentric chromosomes. The centromere is at 14.7 million bp in length. The short arm is rich in tandem repeats [6] and has not been accurately sequenced or annotated. Such sequencing gaps still exist near the centromeres of all chromosomes [3] which

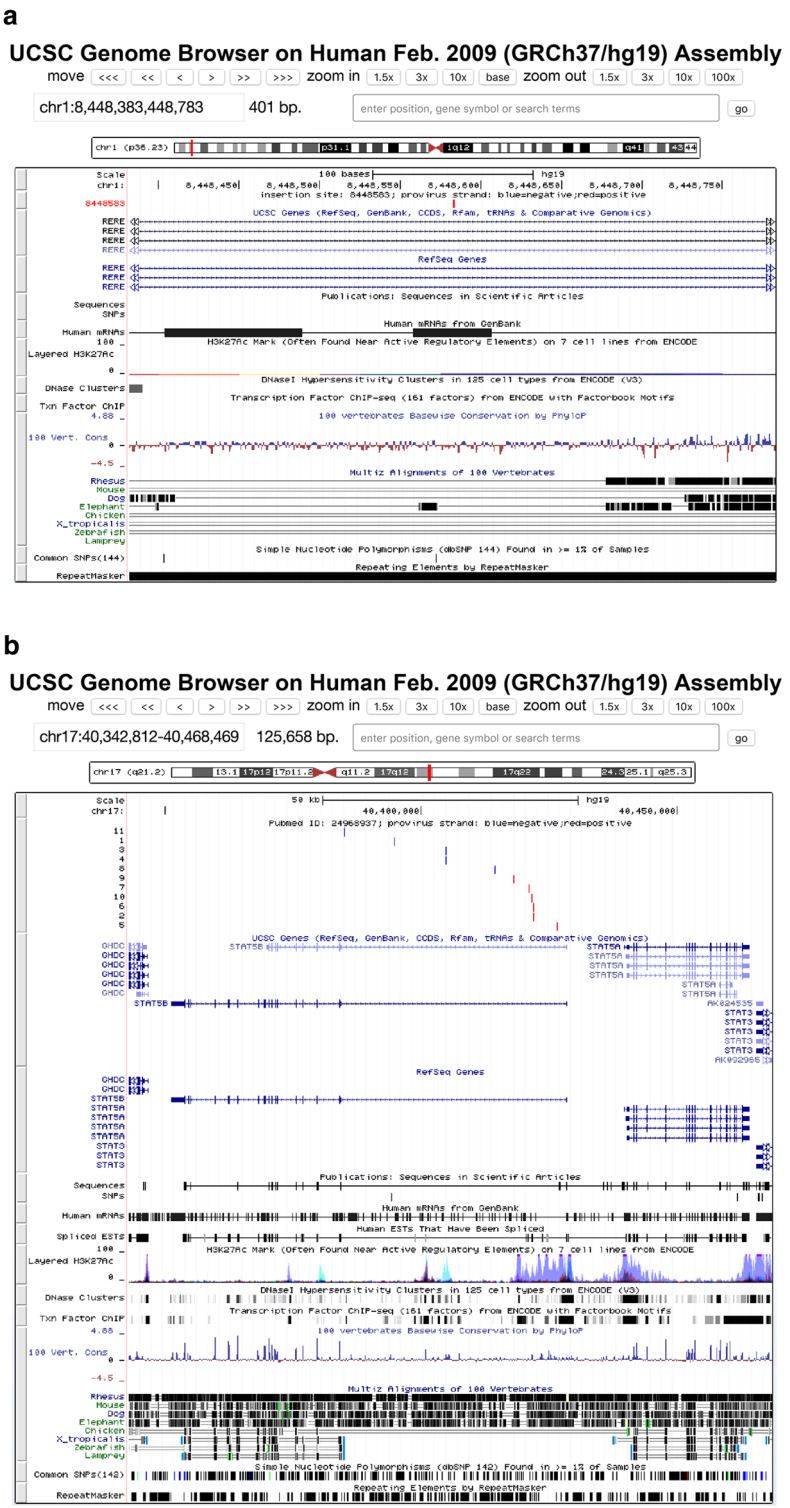

Fig. 4 Integration sites mapped using the UCSC genome browser. *Red vertical bars* show HIV-1 proviruses in the positive orientation relative to the conventional chromosome numbering while *blue vertical bars* show proviruses in the negative orientation. **a** Screen shot from the UCSC genome browser showing the position of an integration site in the RERE gene on human chromosome 1. **b** Screen shot from the UCSC genome browser showing all integration sites in STAT5B gene reported by Maldarelli et al. [12]

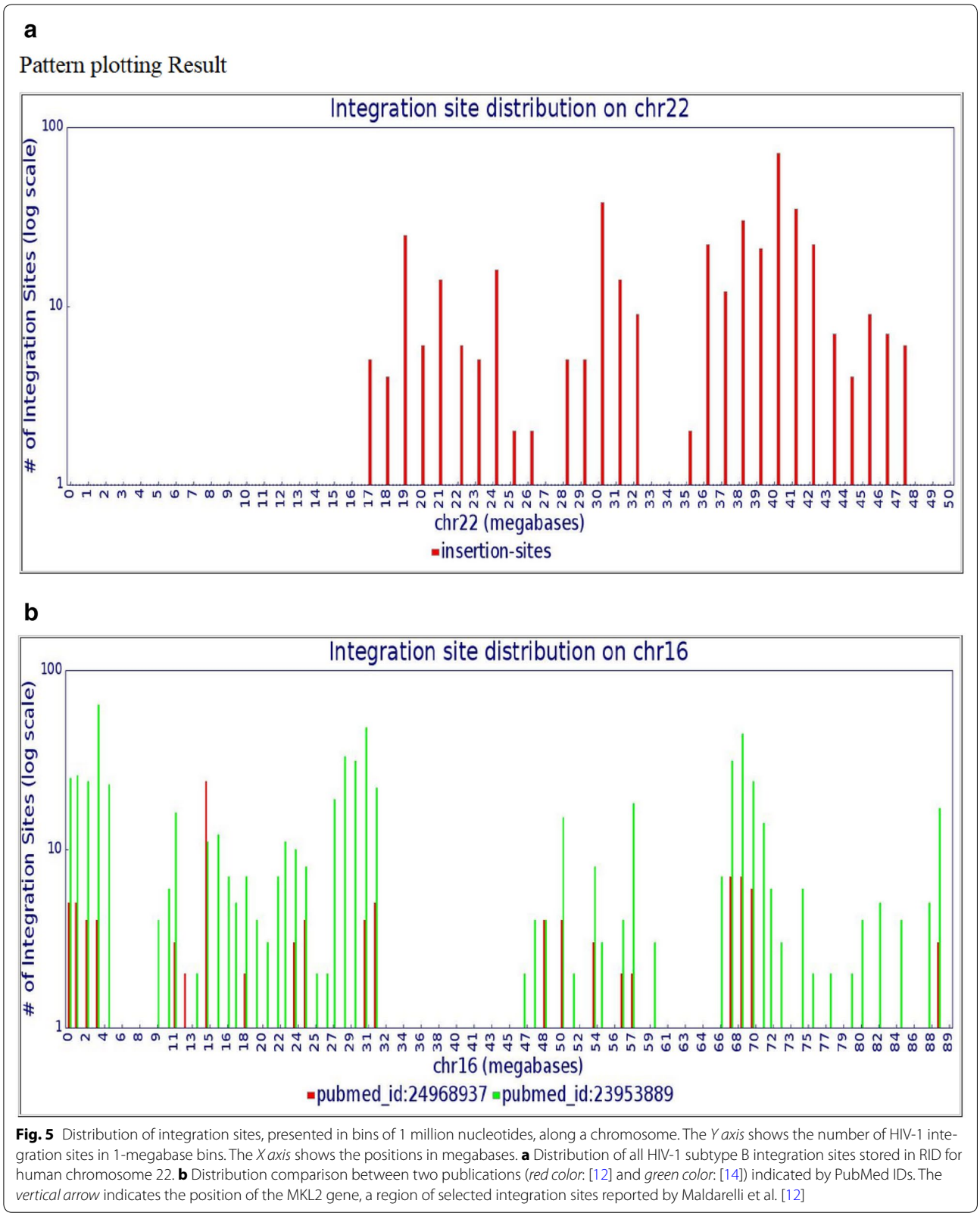

Fig. 5 Distribution of integration sites, presented in bins of 1 million nucleotides, along a chromosome. The *Y axis* shows the number of HIV-1 integration sites in 1-megabase bins. The *X axis* shows the positions in megabases. **a** Distribution of all HIV-1 subtype B integration sites stored in RID for human chromosome 22. **b** Distribution comparison between two publications (*red color*: [12] and *green color*: [14]) indicated by PubMed IDs. The *vertical arrow* indicates the position of the MKL2 gene, a region of selected integration sites reported by Maldarelli et al. [12]

make the discovery of integration sites in these region difficult. The RID tools can also be used with the nucleotide, gene name, PubMed ID, sample, or tissue type selections to display integration site distributions based on these parameters. For example, Fig. 5b shows the comparison of integration site distribution patterns on chromosome 16 from two studies [12, 18].

Uploading data to the database

Users are encouraged to submit their published data to RID. The detailed submission instruction and templates can be accessed in Data Submission tab (Fig. 1). Generally speaking, only data from published peer-reviewed studies will be accepted and made available on the website. We reserve the right not to post data if inspection of the submitted data shows that there are obvious problems with the dataset. In that case, we would contact the authors for clarification.

Conclusion

We have built a large scale, robust relational database called the Retroviral Integration Database (RID) which will be used to store publically available retrovirus integration site data. Users can query all available integration sites or specifically analyze integration sites in specific chromosomes, genes, tissues, etc. Several useful tools are built into the website that are designed to help map integration sites to the UCSC genome browser, to plot integration sites on particular chromosomes, and to determine the flanking host sequences. This database can be used to facilitate meta-analyses of retrovirus integration sites and their chromosomal distribution.

Authors' contributions

WS initiated, designed the database, web interface, and wrote scripts to construct the database and analysis tools. JS designed the database, and designed and wrote scripts to construct the web interface to interact with the database. MK, XW, FM, JWM, BL, JMC, and SHH contributed to the design of the database. All authors read and approved the final manuscript.

Authors' information

JMC was a Research Professor of the American Cancer Society.

Author details

[1] Advanced Biomedical Computing Center, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research (FNLCR), Frederick, MD, USA. [2] HIV Dynamics and Replication Program, NCI, Frederick, MD, USA. [3] Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research (FNLCR), Frederick, MD, USA. [4] Division of Infectious Disease, University of Pittsburgh, Pittsburgh, PA, USA. [5] Department of Molecular Biology and Microbiology, Tufts University, Boston, MA, USA.

Acknowledgements

The authors thank Anne Arthur for adding RID to the NCI HIV DRP web page (http://home.ncifcrf.gov/hivdrp/resources.html), Jon Spindler, David Wells, Shawn Hill, Valarie Boltz, Ann Wiegand, and Uma Mudunuri for their valuable discussions and advice. The authors thank Lucy B. Cook and C. R. Bangham for providing HTLV-1 integration sites and Matthew C. LaFave, M. Burgess for providing MLV integration sites. We thank Connie Kinna and Valerie Turnquist for administrative support.

Competing interests

The authors declare that they have no competing interests.

Funding

We acknowledge the funding sources for this study from NCI CCR, the Office of AIDS Research, NIH, and NCI Contract No. HHSN261200800001E.

Received: 20 May 2016 Accepted: 17 June 2016
Published online: 04 July 2016

References

1. Biasco L, Baricordi C, Aiuti A. Retroviral integrations in gene therapy trials. Mol Ther. 2012;20:709–16.
2. Coffin JM, Hughes SH, Varmus HE. Retroviruses. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1997.
3. Cole CG, McCann OT, Collins JE, Oliver K, Willey D, Gribble SM, Yang F, McLaren K, Rogers J, Ning Z, Beare DM, Dunham I. Finishing the finished human chromosome 22 sequence. Genome Biol. 2008;9:R78.
4. Cook LB, Melamed A, Niederer H, Valganon M, Laydon D, Foroni L, Taylor GP, Matsuoka M, Bangham CR. The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma. Blood. 2014;123:3925–31.
5. De Ravin SS, Su L, Theobald N, Choi U, Macpherson JL, Poidinger M, Symonds G, Pond SM, Ferris AL, Hughes SH, Malech HL, Wu X. Enhancers are major targets for murine leukemia virus vector integration. J Virol. 2014;88:4504–13.
6. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, Ainscough R, Almeida JP, Babbage A, Bagguley C, Bailey J, Barlow K, Bates KN, Beasley O, Bird CP, Blakey S, Bridgeman AM, Buck D, Burgess J, Burrill WD, O'Brien KP, et al. The DNA sequence of human chromosome 22. Nature. 1999;402:489–95.
7. Han Y, Lassen K, Monie D, Sedaghat AR, Shimoji S, Liu X, Pierson TC, Margolick JB, Siliciano RF, Siliciano JD. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. J Virol. 2004;78:6122–33.
8. Hughes SH, Shank PR, Spector DH, Kung HJ, Bishop JM, Varmus HE, Vogt PK, Breitman ML. Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. Cell. 1978;15:1397–410.
9. Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. J Infect Dis. 2007;195:716–25.
10. LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. 2014;42:4257–69.
11. Mack KD, Jin X, Yu S, Wei R, Kapp L, Green C, Herndier B, Abbey NW, Elbaggari A, Liu Y, McGrath MS. HIV insertions within and proximal to host cell genes are a common finding in tissues containing high levels of HIV DNA and macrophage-associated p24 antigen expression. J Acquir Immune Defic Syndr. 2003;33:308–20.
12. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, Coffin JM, Hughes SH. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science. 2014;345:179–83.
13. Serrao E, Engelman AN. Sites of retroviral DNA integration: from basic research to clinical applications. Crit Rev Biochem Mol Biol. 2016;51:26–42.
14. Sherrill-Mix S, Lewinski MK, Famiglietti M, Bosque A, Malani N, Ocwieja KE, Berry CC, Looney D, Shan L, Agosto LM, Pace MJ, Siliciano RF, O'Doherty U, Guatelli J, Planelles V, Bushman FD. HIV latency and integration site placement in five cell-based models. Retrovirology. 2013;10:90.
15. Shin MS, Fredrickson TN, Hartley JW, Suzuki T, Akagi K, Morse HC 3rd. High-throughput retroviral tagging for identification of genes involved in initiation and progression of mouse splenic marginal zone lymphomas. Cancer Res. 2004;64:4419–27.
16. Singh PK, Plumb MR, Ferris AL, Iben JR, Wu X, Fadel HJ, Luke BT, Esnault C, Poeschla EM, Hughes SH, Kvaratskhelia M, Levin HL. LEDGF/p75 interacts

with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. Genes Dev. 2015;29:2287–97.
17. Sunshine S, Kirchner R, Amr SS, Mansur L, Shakhbatyan R, Kim M, Bosque A, Siliciano RF, Planelles V, Hofmann O, Ho Sui S, Li JZ. HIV integration site analysis of cellular models of HIV latency with a probe-enriched next-generation sequencing assay. J Virol. 2016;90:4511–9.
18. Wagner TA, McLaughlin S, Garg K, Cheung CY, Larsen BB, Styrchak S, Huang HC, Edlefsen PT, Mullins JI, Frenkel LM. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science. 2014;345:570–3.
19. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res. 2007;17:1186–94.
20. Wang H, Jurado KA, Wu X, Shun MC, Li X, Ferris AL, Smith SJ, Patel PA, Fuchs JR, Cherepanov P, Kvaratskhelia M, Hughes SH, Engelman A. HRP2 determines the efficiency and specificity of HIV-1 integration in LEDGF/p75 knockout cells but does not contribute to the antiviral activity of a potent LEDGF/p75-binding site integrase inhibitor. Nucleic Acids Res. 2012;40:11518–30.