Bioinformatics, 35(6), 2019, 1073–1075
doi: 10.1093/bioinformatics/bty712
Advance Access Publication Date: 28 August 2018
Applications Note

OXFORD

Databases and ontologies

ISDB: a database toolkit for storing and analyzing viral integration site data

Thomas R. Sibley [1,*], Evan J. Silberman [1,*] and James I. Mullins[1,2]

[1]Department of Microbiology and [2]Departments of Laboratory Medicine, Global Health, and Medicine, University of Washington, Seattle, WA 98195-8070, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on March 2, 2018; revised on May 31, 2018; editorial decision on August 14, 2018; accepted on August 23, 2018

Abstract

Summary: We introduce ISDB, a set of software tools for the creation and administration of relational databases of viral integration site (IS) data. Using ISDB, investigators can curate a private database from any heterogeneous set of data sources, including previously-published datasets and internal, work-in-progress data. To make data visible and accessible to collaborators with varying degrees of computational expertise, ISDB automatically generates web sites describing database contents and data exports in several common formats. Compared to a public depository database, the ability to build local, private databases makes ISDB suitable for use in testing hypotheses and developing analyses in the long pre-publication phase of most research.
Availability and implementation: Installation and usage documentation for ISDB are provided on our website https://mullinslab.microbiol.washington.edu/isdb/. Source code is available under the open source MIT license from https://github.com/MullinsLab/ISDB.
Contact: trsibley@uw.edu or silby@uw.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

While curated datasets of HIV-1 viral genome integration sites exist, we wanted to curate multiple datasets into our own local databases. Our primary goal was a robust database capable of supporting our research from hypothesis to publication, with output that made the data comfortably usable by both software developers and academic researchers with limited database experience. We evaluated the two related projects we could find, but both came up short of our goals:

Chado (Mungall and Emmert, 2007) came with too much overhead and complexity for the benefits it offered. We reasoned it would remain a viable option for the future if we grew out of something else. The Retrovirus Integration Database (RID) (Shao *et al.*, 2016) is designed as a public depository and provides no support for unpublished research data. It is also not open source. Finding neither suitable, we set out building a one-off database for ourselves, and over time we separated the toolkit we created from the data itself. The toolkit became ISDB.

2 Design

ISDB comprises three core components: a simple-to-use relational database schema, documented conventions for describing diverse datasets of integration sites, and a set of command-line and web-based tools for working with the schema and datasets.

Organizing integration site data with ISDB allows researchers to store data in a robust relational database system, and thereby easily use other ISDB tools to produce:

- An easy-to-use website describing the data, linking to bulk downloads, data summaries and documentation
- Ready-made exports for UCSC Genome Browser, IGV, Geneious and more genome browsing and analysis tools
- Frozen versions of the relational database and website in order to lock analyses to an unchanging dataset
- A set of curated data sources representing the raw, uncollated data, with preparation history and provenance tracked

2.1 Data sources

A guiding principle of ISDB's architecture is that the derived database contents are not as precious as the raw data. Updates for a given data source are performed by deleting and reloading all data from that source. This side-steps complicated issues around applying partial updates to previously loaded data, and the handling of source data and database management is greatly simplified by most tasks being idempotent. Thus, by design, a given database can be quickly reproduced from the collection of data sources that initially went into it.

An ISDB source is simply a directory of files that follows a few conventions. Sources are responsible for transforming the relevant experimental data into a file of comma-separated values (CSV) which conforms to certain requirements. The amount of work a source has to do to transform the starting data into the expected CSV file will vary greatly. Some sources may only require a minimal transformation of an existing tabular data file of predetermined integration sites. Others may start with primary sequencing data, automatically map the integration sites, and output an appropriate CSV. ISDB does not dictate how to generate data, only that the data meets the documented minimal expectations.

Sources are ideally kept in version-controlled repositories where their individual history is tracked. By tracking data sources separately from the database itself, the transformation of original data is traceable, verifiable and reproducible.

2.2 Broad accessibility curve

To make data visible and accessible to a range of skill levels, from bench scientists to bioinformaticians to software developers, ISDB provides many means of access.

The foundational data access method is through direct SQL queries, which expose the full power of a relational database. SQL queries are the *lingua franca* of relational databases and allow expressive filtering, summarizing and comparison of datasets. Queries may be run from any programming language or environment, such as R, Python, Perl, as well as interactively using psql. Software packages like R's dplyr can also abstract away some of the syntax of SQL into more familiar R operations.

However, many tasks on the data can be accomplished without knowing SQL by using data files exported by ISDB tools. These files can be exported on demand or configured for automatic updates on a regular schedule. ISDB exports three tabular datasets by default to meet a variety of basic analysis needs:

- A summary of host genes containing integration sites, with counts
- A summary of integration sites by subject, with multiplicity calculated
- The previous, with the containing genes annotated

Each dataset is available as a CSV file, an Excel spreadsheet and a JSON file. Providing a direct Excel export, rather than requiring import from CSV, avoids common file conversion issues [Such as Excel treating gene names like 'SEPT2' as the date September 2nd (Zeeberg *et al.*, 2004; Ziemann *et al.*, 2016).] and reduces the steps required for analysts to use the data.

Genome browser tracks in BED format are produced for loading into the UCSC Genome Browser, the Integrated Genomics Viewer (IGV) and many other software tools. BED files are suitable for both visualizing the integration sites and performing calculations on them using bedtools.

For focusing on a particular gene, a GFF3 file is exported for each gene containing integration sites. The coordinates in the GFF3 file are relative to the gene itself, which makes the files suitable for annotating a gene sequence, such as those downloadable from NCBI Gene.

2.3 Frozen versions

ISDB can freeze a copy of a database at a point in time, alongside the generated website and all data exports. This allows analysis to be done on a fixed point and easily reproduced since the analyzed data will not drift as data accumulation proceeds: researchers can prepare analyses for publication using the frozen data while others continue to generate new data.

2.4 Extensible metadata

In addition to a basic set of fields, ISDB also provides a standardized way to include additional, optional information per IS. Extensible metadata is critical for cross-linking integration site data to related records in a variety of other databases, such as GenBank, PubMed and laboratory information management systems.

2.5 Gene overlaps are computed, not stored

ISDB stores the locations of host genes and the locations of integration sites separately. Overlapping genes are computed on-the-fly when making a query. Parameters, such as the maximum distance allowed for 'nearby genes', can be adjusted according to the needs of the researcher. Compared to determining annotation overlaps once during data loading, our approach is more flexible and provides better data integrity by decoupling the gene database from the table of integration sites. For example, it is impossible within ISDB to make the otherwise easy mistake of associating one IS dataset with version *A* of NCBI gene names and another IS dataset with version *B* of NCBI gene names. It's also trivial to update the NCBI Gene annotation data to a more recent release since it does not require modifying the integration site data.

3 Example

As a public example of a database built using ISDB, we provide HIRIS (https://mullinslab.microbiol.washington.edu/hiris/), which aggregates data from studies of HIV-1 site selection *in vivo* from patients on suppressive antiretroviral therapy (Han *et al.*, 2004; Ikeda *et al.*, 2007; Mack *et al.*, 2003; Maldarelli *et al.*, 2014; Wagner *et al.*, 2014) and from *in vitro* models of latency (Lewinski *et al.*, 2005; Pace *et al.*, 2012; Shan *et al.*, 2011; Sherrill-Mix *et al.*, 2013; Sunshine *et al.*, 2016; Wang *et al.*, 2007). Thanks to the design of ISDB, we are able to maintain a private version of HIRIS for our pre-publication research and make data available in the public version as we publish on it.

A guide to getting started with ISDB is available in the Supplementary material. Complete documentation is available online at https://mullinslab.microbiol.washington.edu/isdb/doc/.

Funding

This work was supported by the National Institutes of Health to J.I.M. [R01AI111806, R21AI122361]; and the Functional Profiling and Computational Biology Core of the University of Washington's Center for AIDS Research [P30 AI027757].

Conflict of Interest: none declared.

References

Han,Y. *et al.* (2004) 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.*, **78**, 6122–6133.

Ikeda,T. *et al*. (2007) Recurrent HIV-1 integration at the bach2 locus in resting cd4+ t cell populations during effective highly active antiretroviral therapy. *J. Infect. Dis*., **195**, 716–725.

Lewinski,M.K. *et al*. (2005) Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol*., **79**, 6610–6619.

Mack,K.D. *et al*. (2003) 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. *Journal of Acquir. Immune Defic. Syndr*., **33**, 308–320.

Maldarelli,F. *et al*. (2014) HIV latency. specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*, **345**, 179–183.

Mungall,C.J. and Emmert,D.B. (2007) A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337.

Pace,M.J. *et al*. (2012) Directly infected resting cd4+t cells can produce HIV gag without spreading infection in a model of HIV latency. *PLoS Pathogens*, **8**, e1002818.

Shan,L. *et al*. (2011) Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. *J. Virol*., **85**, 5384–5393.

Shao,W. *et al*. (2016) Retrovirus integration database (rid): a public database for retroviral insertion sites into host genomes. *Retrovirology*, **13**, 47.

Sherrill-Mix,S. *et al*. (2013) HIV latency and integration site placement in five cell-based models. *Retrovirology*, **10**, 90.

Sunshine,S. *et al*. (2016) HIV integration site analysis of cellular models of HIV latency with a probe-enriched next-generation sequencing assay. *J. Virol*., **90**, 4511.

Wagner,T.A. *et al*. (2014) HIV latency. proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*, **345**, 570–573.

Wang,G.P. *et al*. (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res*., **17**, 1186–1194.

Zeeberg,B.R. *et al*. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using excel in bioinformatics. *BMC Bioinformatics*, **5**, 80.

Ziemann,M. *et al*. (2016) Gene name errors are widespread in the scientific literature. *Genome Biol*., **17**, 177.