Reid *et al. BMC Res Notes (2020) 13:195*
https://doi.org/10.1186/s13104-020-05038-w

BMC Research Notes

RESEARCH NOTE

Open Access

Automated gene data integration with Databio

Robert W. Reid[1,2], Jacob W. Ferrier[1] and Jeremy J. Jay[1,2*]

Abstract

Objective: Although sequencing and other high-throughput data production technologies are increasingly affordable, data analysis and interpretation remains a significant factor in the cost of -omics studies. Despite the broad acceptance of findable, accessible, interoperable, and reusable (FAIR) data principles which focus on data discoverability and annotation, data integration remains a significant bottleneck in linking prior work in order to better understand novel research. Relevant and timely information discovery is difficult for increasingly multi-disciplinary projects when scientists cannot easily keep up with work across multiple fields. Computational tools are necessary to accurately describe data contents, and empower linkage to existing resources without prior knowledge of the various database resources.

Results: We developed the Databio tool, accessible at https://datab.io/, to automate data parsing, identifier detection, and streamline common tasks to provide a point-and-click approach to data manipulation and integration in life sciences research and translational medicine. Databio uses fast real-time data structures and a data warehouse of 137 million identifiers, with automated heuristics to describe data provenance without highly specialized knowledge or bioinformatics training.

Keywords: Data integration, Workflow automation, Knowledge discovery

Introduction

Although sequencing and other high-throughput data production technologies are increasingly affordable, data analysis remains a significant factor in the cost of -omics studies [1]. Without improving the ability to automate data integration and interoperation, the cost of analysis will continue to impede access to precision medicine for underserved populations with limited resources. Many resources have been developed around the concept of a central “Data Commons”, but the path forward remains unclear [2], and current large data repositories are highly specialized and difficult to apply broadly. Despite the acceptance and proliferation the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles [3], current data provider implementations focus on descriptive metadata and keyword-oriented search applications, leaving the detailed gene and other -omics data inaccessible to computational discovery methods.

Data producers recognize the need to enable greater access to hosted data, but there are no well-accepted machine-readable means for annotating the contents of data sets across the biomedical landscape [4]. The lack of available standards and tools make it a cumbersome and time-consuming task to properly annotate identifier sources, record their provenance throughout an analytical process, and track subsequent data quality metrics. These challenges exist regardless of the level of research activity, including mammalian, marine, and agricultural research domains [5–7]. As a result, the majority of useful scientific results remain buried in supplementary tables, figures, and poorly indexed data archives.

*Correspondence: jeremy.jay@uncc.edu
[1] Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA
Full list of author information is available at the end of the article

Although manual curation efforts have led to increasingly more data becoming available in data portals and publication annotations, these efforts require specialized knowledge around biomedical resources. Even seemingly trivial tasks are burdensome, such as those required for secondary analysis of a gene list in a supplementary table. One must be able to identify obscure identifiers such as 'ENSG00000168653', identify tools or mapping data that support it, and translate into symbols (e.g. 'NDUFS5') or identifers (Entrez Gene ID 4725, or RefSeq Accession NM_004552.3, etc) useful for their own analysis methods. Using these resources necessitates experience with the extract-transform-load (ETL) process, and the resource knowledge and technical expertise has little to do with the science itself.

These challenges represent an increasing burden on data producers, which is deferred to data consumers who are faced with the need to integrate loosely described high-throughput experiments into novel studies [8]. Because data consumers only need these analytical skills occasionally, they are more prone to implementation errors and struggle to fully integrate complex data relationships [9, 10]. Thus there is a need to simplify and automate the discovery and retrieval process.

Main text

We present Databio, a novel framework for automating the extraction, annotation, and integration of gene-oriented data sets. Databio automates data parsing and identifier detection, and streamlines many common tasks to provide a point-and-click approach to data manipulation and integration across a broad spectrum of applications in life sciences research and translational medicine. This ability to quickly and accurately streamline complex tasks will enable faster and better analysis of -omics data.

Implementation and available data

Databio is implemented as a web-based data portal (https://datab.io) that allows users to interact with the embedded tools using an interactive web browser-based interface.

User data uploads are first handled via an automatic detection framework that determines the source data format (see top Fig. 1). The current implementation supports Tab-separated values (TSV), Comma-separated values (CSV), and Excel 2007+ spreadsheets (XLSX). Records (rows) and fields (columns) within these documents are exposed to the rest of the application through a modular interface allowing for support for more data formats in future software updates. Heuristic techniques are applied to the parsed data to remove headers and determine field labels, allowing for a more descriptive display interface (see Fig. 1).

Once fields are parsed, values are aggregated together and searched against our warehouse of multiple gene identifier data sources. Our current snapshot contains over 137 million unique gene, transcript, and protein identifiers and 92 million unique mapping pairs (Table 1). Despite the extreme scale of determining identifier source, this classification can be completed accurately in real-time (less than 1 s) using Bloom filters for fast approximate matching [11]. The top hits for each field are collected (along with sample values) and returned to the web interface so that users can verify the accuracy of the predicted identifier type.

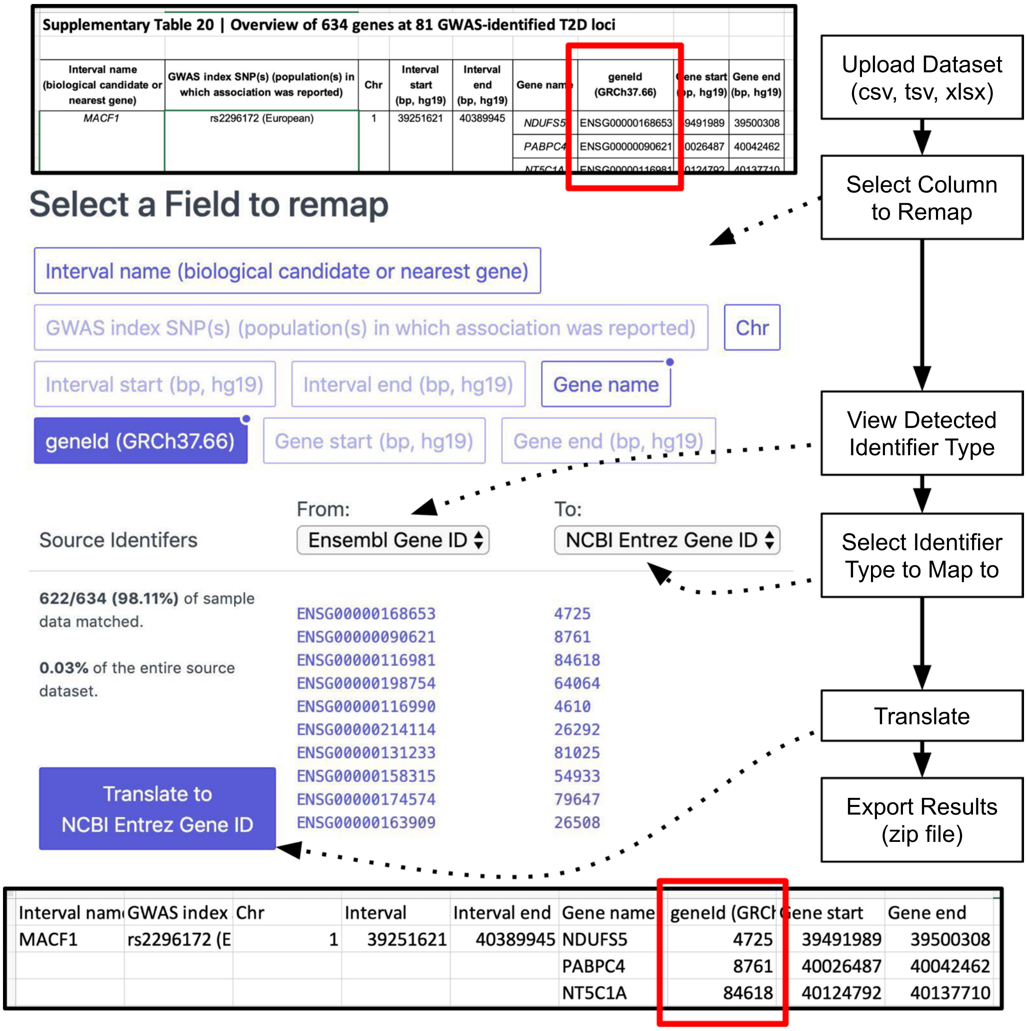
In addition to the classification index representation, the Databio database also contains mappings that allow supported identifiers to be translated into other identifier types. Although this common task has been supported by other tools such as David, Uniprot, and BioMart [12–14], these tools require manual data manipulation, specialized knowledge of identifier sources, and cannot replace identifiers within the context of the original data file [15]. Databio is able to translate identifiers in-place, removing multiple opportunities for error and keeping the data in context. These changes are applied to the existing data schema and exported to a CSV-format data set that can be readily imported into other tools for subsequent analysis (see bottom of Fig. 1).

Further easing the burden of data manipulation on the user, Databio is able to track important data quality issues such as missing identifiers and ambiguous mappings. The Databio warehouse maintains a record of publication and citation info for each identifier source, the last fetch and access dates, and analysis logs describing processing steps and data quality metrics. Using this information, Databio can establish that necessary metadata for publication, distribution, and reuse is present and accurately tracked. This ensures that data consumers know the state of a data set including access dates, citations, and relevant usage limitations.

Usage

For example, a study identified 634 genes associated with Type 2 Diabetes Genome-Wide Association Study loci [16], and provided the results in a Supplementary Table (see top, Fig. 1). We want to look for relationships between the RefSeq Transcript sequences of the genes and the listed loci. However, searching for 'ENSG00000168653' in RefSeq currently yields no results, and the gene Symbol 'NDUFS5' returns 19 Human results. One must translate the gene identifiers into more specific RefSeq Transcript IDs.

Upon visiting the Databio site, the user is able to upload this Excel file (or any other TSV, CSV or XLSX data file) even though it does not fit a pre-determined field



Supplementary Table 20 | Overview of 634 genes at 81 GWAS-identified T2D loci

Interval name (biological candidate or nearest gene)	GWAS index SNP(s) (population(s) in which association was reported)	Chr	Interval start (bp, hg19)	Interval end (bp, hg19)	Gene nam	geneId (GRCh37.66)	ene start p, hg19)	Gene end (bp, hg19)
MACF1	rs2296172 (European)	1	39251621	40389945	*NDUFS5*	ENSG00000168653	9491989	39500308
					PABPC4	ENSG00000090621	0026487	40042462
					NT5C1A	ENSG00000116981	0124792	40137710

Select a Field to remap

Interval name (biological candidate or nearest gene)
GWAS index SNP(s) (population(s) in which association was reported)
Chr
Interval start (bp, hg19)
Interval end (bp, hg19)
Gene name
geneId (GRCh37.66)
Gene start (bp, hg19)
Gene end (bp, hg19)
Upload Dataset (csv, tsv, xlsx)
Select Column to Remap
View Detected Identifier Type
Select Identifier Type to Map to
Translate
Export Results (zip file)
Translate to NCBI Entrez Gene ID

Source Identifers	From: Ensembl Gene ID	To: NCBI Entrez Gene ID
622/634 (98.11%) of sample data matched.	ENSG00000168653	4725
	ENSG00000090621	8761
0.03% of the entire source dataset.	ENSG00000116981	84618
	ENSG00000198754	64064
	ENSG00000116990	4610
	ENSG00000214114	26292
	ENSG00000131233	81025
	ENSG00000158315	54933
	ENSG00000174574	79647
	ENSG00000163909	26508

Interval nam	GWAS index	Chr	Interval	Interval end	Gene name	geneId (GRCI	ene start	Gene end
MACF1	rs2296172 (E	1	39251621	40389945	NDUFS5	4725	39491989	39500308
					PABPC4	8761	40026487	40042462
					NT5C1A	84618	40124792	40137710

Fig. 1 Databio web interface workflow showing data upload (including Excel formatting, headers, and merged fields). Point-and-click field mapping allows selection of source and replacement gene identifiers. Results are then exported with new identifiers. Statistics, bibliography, and provenance files are included in download archive but not shown

Table 1 Gene identifier sources loaded into Databio as of 2019-09-12

Name	Subsets	Total	References
NCBI Entrez Gene	39	25,295,958	[17]
RefSeq Transcripts	1	2,211,841	[18]
RefSeq Proteins	1	40,574,328	[18]
Ensembl Gene	207	5,442,203	[19]
Ensembl Transcripts	207	9,000,822	[19]
Ensembl Proteins	207	6,923,465	[19]
KEGG Genes	6128	29,541,384	[20]
UniprotKB/Swiss-Prot	1	18,493,595	[13]
HGNC Gene IDs	1	42,050	[21]
HGNC Symbol	1	42,050	[21]
HGNC Gene Names	1	42,050	[21]
OMIM Genes	1	16,197	[22]

layout. Column names (fields) are automatically parsed and identified for selection on the second page (see top, Fig. 1). Fields with high-quality automated classification are marked with a circle in the top right corner to indicate a high correspondence to a known Databio identifier source (For example, the blue box "geneId (GRCh37.66)" in Fig. 1). The user is then able to click on the field name that they want to remap. The exact match rate, as well as the percent coverage of the corresponding source dataset, is shown to the user under the 'Source Identifiers' header on the left.

We can see that for this example, even though the file did not explicitly mention the source of gene identifiers, Databio easily determined them to be Ensembl Gene IDs. For other data sets, if there is more ambiguity to the identifiers (e.g. integers), the user can use the drop-down on the left to see the other matched identifiers sources and find the most appropriate choice. The user can then choose the desired identifier type to map to, using the drop-down on the right, and an automatically generated list of identifiers that map to the original identifier source. Changing either the 'to' or 'from' drop-down selections automatically updates to display a sample of the original identifiers from the uploaded data, and the associated remapped identifiers so that the user can confirm expectations. Finally, the user may begin the translation processing, which leads to a new page including the remapped data file for download, statistics, some text describing the methods and data sources used with a bibliography and analysis logs. This information is all available in a compressed ZIP archive ensuring that important information is delivered together as one unit.

Discussion

Databio automates and streamlines the process of gene identifier translation, enabling new approaches to data-driven discovery by lowering the burden of data manipulation and prior knowledge of biomedical resources. Support for more identifier sources, more data formats, and chained identifier conversions (A → B → C) will greatly increase the utility of Databio across the life sciences. In addition, future computational analyses will build upon this base, enabling data set search based on related data contents and not just shared metadata. Together these improvements will enable future machine learning applications by removing the need for manual intervention in data import processes, shortening learning times and improving the pace of data-driven discovery.

Limitations

- Primarily gene-centric automated identifier detection. We are working to expand the data warehouse to include other data types. These methods will require further work to allow identification in the presence of noise or natural language (e.g. clinical reports).
- Cannot handle chained/multi-step conversions. e.g. to translate from A to X if there is no direct mapping, manual translation to an intermediate value is necessary first (A to B, then B to X). This is likely unintuitive to new users but an issue we hope to address in the future.
- Search methods currently scale linearly with search scope. e.g. as the data warehouse grows, so does the search time. We are working on algorithmic methods and data structures to address this limitation.

Abbreviations

FAIR: Findable, accessible, interoperable, and reusable; ETL: Extract transform load; TSV: Tab-separated values; CSV: Comma-separated values; XLSX: Excel Microsoft Office Open XML Format Spreadsheet; ZIP: A compressed archive format; NCBI: National Center for Biotechnology Information; KEGG: Kyoto Encyclopedia of Genes and Genomes; HGNC: HUGO Gene Nomenclature Committee; HUGO: Human Genome Organisation; OMIM: Online Mendelian Inheritance in Man; GNU: GNU's Not Unix!; GPL: General Public License; XML: eXtensible Markup Language.

Acknowledgements

Not applicable.

Authors' contributions

JJJ conceived the project, determined the analysis methods and materials, and provided overall direction. JJJ and JWF designed and implemented the software. JWF designed the figure. JJJ and RWR wrote the paper. RWR and JWF provided corrections and feedback. All authors read and approved the final manuscript.

Funding

Not applicable.

Availablility of data and materials

A public instance is accessible at https://datab.io/, and uses a data warehouse of identifiers. Source code and binaries for the web tool can be downloaded from https://github.com/joiningdata/databio/. The software is supported on Linux, Windows, and macOS operating systems using the Go programming language.

All data used for this project is publicly accessible (See Table 1). Source code and ETL scripts for the data populating the data warehouse can be found at https://github.com/joiningdata/databio_sources/.

Project name: Databio
Project home page: https://datab.io
Archived version: 0e277ee72f353cd7ed9d0f5ef005f9f14b11618a
Operating system(s): Platform independent
Programming language: Go
License: GNU GPL v3

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA. [2] North Carolina Research Campus, 150 N Research Campus Dr, Kannapolis, NC 28081, USA.

Received: 29 January 2020 Accepted: 20 March 2020
Published online: 01 April 2020

References

1. Mardis ER. The \$1,000 genome, the \$100,000 analysis? Genome Medicine. 2010;2(11):84. https://doi.org/10.1186/gm205.
2. NIH Common Fund: New Models of Data Stewardship—Data Commons Pilot. https://commonfund.nih.gov/commons. Accessed 09 Jan 2020.
3. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
4. National Research Council. Barriers to the use of Databases. In: Pool, R., Esnayra, J. (eds.) Bioinformatics: converting data to knowledge. Washington, DC: The National Academies Press; 2000. https://doi.org/10.17226/9990.

5. Maughan PJ, Lee R, Walstead R, Vickerstaff RJ, Fogarty MC, Brouwer CR, Reid RR, Jay JJ, Bekele WA, Jackson EW, Tinker NA, Langdon T, Schlueter JA, Jellen EN. Genomic insights from the first chromosome-scale assemblies of oat (*Avena* spp.) diploid species. BMC Biol. 2019;17(1):92. https://doi.org/10.1186/s12915-019-0712-y.
6. Janies DA, Witter Z, Linchangco GV, Foltz DW, Miller AK, Kerr AM, Jay J, Reid RW, Wray GA. EchinoDB, an application for comparative transcriptomics of deeply-sampled clades of echinoderms. BMC Bioinf. 2016;17:48. https://doi.org/10.1186/s12859-016-0883-2.
7. Logan RW, Robledo RF, Recla JM, Philip VM, Bubier JA, Jay JJ, Harwood C, Wilcox T, Gatti DM, Bult CJ, Churchill GA, Chesler EJ. High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. Genes Brain Behav. 2013;12(4):424–37. https://doi.org/10.1111/gbb.12029.
8. Bubier JA, Wilcox TD, Jay JJ, Langston MA, Baker EJ, Chesler EJ. Cross-species integrative functional genomics in GeneWeaver reveals a role for Pafah1b1 in altered response to alcohol. Front Behav Neurosci. 2016;10:1. https://doi.org/10.3389/fnbeh.2016.00001.
9. Jay JJ, Chesler EJ. Performing integrative functional genomics analysis in GeneWeaver.org. In: Gene function analysis. Methods in molecular biology, vol. 1101. Totowa: Humana Press; 2014, pp. 13–29. https://doi.org/10.1007/978-1-62703-721-1
10. Jay JJ. Cross species integration of functional genomics experiments. Int Rev Neurobiol. 2012;104:1–24. https://doi.org/10.1016/B978-0-12-398323-7.00001-X.
11. Bloom BH. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM (1970). Accessed 09 Jan 2020
12. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):3.
13. UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):204–12. https://doi.org/10.1093/nar/gku989
14. Smedley D, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):589–98. https://doi.org/10.1093/nar/gkv350.
15. Jay JJ, Sanders A, Reid RW, Brouwer CR. Connecting nutrition composition measures to biomedical research. BMC Res Notes. 2018;11(1):883. https://doi.org/10.1186/s13104-018-3997-y.
16. Fuchsberger C, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–7. https://doi.org/10.1038/nature18642.
17. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2007;35(Database issue):26–31. https://doi.org/10.1093/nar/gkl993
18. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(Database issue):501–4. https://doi.org/10.1093/nar/gki025.
19. Zerbino DR, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):754–61. https://doi.org/10.1093/nar/gkx1098.
20. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36(Database issue):480–4. https://doi.org/10.1093/nar/gkm882.
21. Yates B, Braschi B, Gray KA, Seal Rl, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017;45(D1):619–25. https://doi.org/10.1093/nar/gkw1033.
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):514–7. https://doi.org/10.1093/nar/gki033.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.