Agent-Assisted Metadata Discovery and Validation for Genomic Data-sharing

Sam Blouir
George Mason University
Fairfax, VA, USA
sblouir@gmu.edu

Flavia Negrete*
Broad Institute of MIT and Harvard
Cambridge, MA, USA
FlaviaNegrete@gmail.com

Amarda Shehu
George Mason University
Fairfax, VA, USA
ashehu@gmu.edu

ABSTRACT

AI-driven genomic research remains hindered by fragmented repositories, inconsistent metadata standards, and uncertainties about data validity and reuse. We propose a federated meta-layer to aggregate genomic datasets from multiple public sources into a unified discovery repository. To automate validation and enhance reliability, we introduce a proof-of-concept large language model-based agent that autonomously generates and executes code to validate dataset availability, structural integrity, and labeling correctness. Inspired by autonomous scientific discovery frameworks such as the robot scientist [11], our approach simplifies genomic dataset discovery and verification, aiding the exchange of genomic discovery among different research facilities by enhancing the efficiency of genomic database retrieval and curation. Our proof-of-concept implementation is publicly available at https://github.com/samblouir/agents_for_genomics.

ACM Reference Format:
Sam Blouir, Flavia Negrete, and Amarda Shehu. 2025. Agent-Assisted Metadata Discovery and Validation for Genomic Data-sharing. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Genomic research increasingly leverages AI for tasks such as variant detection and disease prediction. Fragmentation across repositories (NIH dbGaP [1, 8], EGA [13], and individual labs) and inconsistent metadata hinder data discoverability and trustworthiness. Despite standardization efforts (GA4GH [9], ELIXIR [2]), verifying metadata accuracy remains challenging due to manual or nonexistent quality checks [4, 5].

Current genomic data-sharing practices have resulted in widespread fragmentation. Datasets are scattered across public repositories and individual lab websites, each using different metadata standards and documentation practices. Community initiatives like GA4GH [9] and ELIXIR [2] have improved some standardization, yet challenges remain in verifying metadata accuracy and dataset reliability, especially when considering the size of these data sources [4].

*Work done while at the Broad Institute of MIT and Harvard

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Existing public genomic databases provide valuable open-access data, but inconsistencies in metadata standards and limited quality verification pose significant challenges. Labs independently manage metadata annotations, sometimes resulting in stale, mislabeled, or poorly structured data [5]. Community-driven initiatives have partially addressed standardization, but data validity and quality checks remain manual, inconsistent, or, in the worst case, nonexistent.

We propose a lightweight, federated meta-layer that enhances genomic data discoverability while promoting responsible AI governance. Our approach combines the aggregation of multiple sources (including HuggingFace [16], NCBI [3], and other sources of genomic datasets), and introduces an automated AI-agent pipeline that validates dataset structure, labels (if applicable), and metadata accuracy, helping ensure trustworthy data sharing without excessive manual burden from adjusting past projects [6, 7].

Primarily inspired by successful community-oriented platforms such as HuggingFace, we propose a federated meta-layer that integrates automated dataset validation and searching. Our approach seeks not only to enhance data discoverability but also to ensure metadata accuracy and structural integrity through an AI-driven validation agent.

2 TECHNICAL APPROACH: FEDERATED META-LAYER WITH AI VERIFICATION

Our architecture includes two complementary components:

2.1 Search Engine

By spreading and aggregating search results across multiple sources, our meta-layer can allow researchers to more easily learn where data can be directly sourced, find out about additional disclaimers, and learn how to correctly cite or reuse it [6, 7]. PIs and labs can see direct benefits of clarifying their documentation for AI-driven verification. Having verified usage guides can lead to increased citations and collaborations.

2.2 AI Agent for Data Validity and Structure Verification

To proactively ensure data quality and structural accuracy, we deploy an automated AI agent that performs routine dataset validation tasks:

The AI agent can attempt to load and parse dataset samples, ensuring they match advertised metadata structures. For example, documentation should be straightforward to follow. Clear discrepancies can be discovered by an agent, reducing code-related overhead and annoyances involved. The agent writes code to train a small

classifier (or other model) on this data, and make basic verifications that elements such as sequence labels are aligned. Possible annotation errors can be automatically flagged for review based on excessive model non-convergence.

2.3 Proof-of-Concept Agent Implementation

We demonstrate a working proof-of-concept by leveraging NousResearch Mistral 24B DeepHermes [15] as an agent. We provide our code at https://github.com/samblouir/agents_for_genomics. Here, the agent locates the dataset online, downloads it, and attempts to validate the genomic labels by training a small model on the "human vs. worm" dataset from GenomicsBenchmarks [10]. We note this is a simple and minimal example. Passing automated checks like these, like other forms of software testing, can help improve trust in metadata and labels.

3 USE CASE: ANTIBIOTIC-RESISTANCE DATASET

A researcher studies antibiotic-resistant genes in bacterial pathogens. Previously, this required manual data discovery across isolated websites with uncertain metadata accuracy.

With our meta-layer and AI agent:

(1) A researcher can search and receive aggregated results from NIH, independent lab pages, and other sources [7, 9].
(2) Each entry includes validated metadata, direct links, explicit usage guidelines, and ready-to-run AI code examples for immediate integration into their research.
(3) Automated AI-agent verification can help assure the researcher of metadata accuracy.

4 EVALUATION

Quantitative measures of dataset discovery can be clearly marked by watching historical download and citation rates for relevant papers. Model outputs can be audited by authors or other community members verifying and contesting flagged datasets.

5 DISCUSSION

A meta-layer with AI-driven validation can significantly enhance genomic data discovery and responsible usage by centralizing methods to communicate with authors, distribute datasets, and share documentation and code to use the data. Such an automated platform can promote responsible, efficient data reuse with potentially minimal manual overhead.

Although platforms such as HuggingFace have gained wide popularity in the NLP and computer vision domains, its adoption in the genomics community remains limited [14]. Challenges include integrating vastly different data formats, maintaining accurate metadata and instructions for large (and often sensitive) datasets, and addressing other considerations unique to genomics. Community-led governance can help establish standards for an approach that could allow researchers to benefit from the speed and ease of a platform similar to HuggingFace [12].

Additionally, continuous refinement or upgrades to the AI-driven verification methods may be needed to keep pace with the evolving field—but we have seen standards naturally emerge in other fields as incentives exist, such as increased citations.

6 CONCLUSION

We propose a community-led meta-layer website and dataset congregator to help boost the proliferation and implementation of genomic data for researchers, augmented by an AI-driven verification agent that attempts to verify metadata validity and structure. With enhanced verification techniques, such as automated BLAST sequence checking and improved data integrity verification, such a solution can promote trustworthy and transparent genomic data sharing. This platform simplifies responsible AI reuse and reduces barriers for researchers, advancing efficient and ethical genomic AI research.

7 LIMITATIONS

Although our proof-of-concept agent works to verify a relatively simple binary genomic classification task, real-world datasets can have more complex structures or require domain-specific validation steps. Privacy, ethical guidelines, and regulatory considerations for human genomic data also demand careful planning and might limit data sharing distribution or even links or references to the source, in certain contexts. Future work should address these aspects to ensure that AI-assisted validation can be robust and inclusive to researchers throughout the field.

REFERENCES

[1] [n. d.]. Database of Genotypes and Phenotypes (dbGaP). https://www.ncbi.nlm.nih.gov/gap/.
[2] [n. d.]. ELIXIR: Accelerating Data-Driven Research in Europe. https://elixir-europe.org/.
[3] [n. d.]. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/.
[4] Anna Bernasconi, Arif Canakoglu, Marco Masseroli, and Stefano Ceri. 2022. META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 1 (2022), 543–557. https://doi.org/10.1109/TCBB.2020.2998954
[5] Aylin Caliskan, Seema Dangwal, and Thomas Dandekar. 2023. Metadata integrity in bioinformatics: Bridging the gap between data and knowledge. *Computational and Structural Biotechnology Journal* 21 (2023), 4895–4913. https://doi.org/10.1016/j.csbj.2023.10.006
[6] Arif Canakoglu, Anna Bernasconi, Andrea Colombo, Marco Masseroli, and Stefano Ceri. 2019. GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database (Oxford)* 2019 (2019), baz132. https://doi.org/10.1093/database/baz132
[7] Xiaoling Chen, Anupama E. Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, Yueling Li, Nansu Zong, Min Jiang, Deevakar Rogith, Mandana Salimi, Hyeon-Eui Kim, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Claudiu Farcas, Todd Johnson, Ron Margolis, George Alter, Susanna-Assunta Sansone, Ian M. Fore, Lucila Ohno-Machado, Jeffrey S. Grethe, and Hua Xu. 2018. DataMed – an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association* 25, 3 (2018), 300–308. https://doi.org/10.1093/jamia/ocx121
[8] Scott Federhen, Karen Clark, Tanya Barrett, Tanya Diekmann, Lewis Y Geer, Dustin Gonzales, Marietta Hebert, Ilene Karsch-Mizrachi, Avi Kimchi, Kim D Pruitt, and et al. 2023. The NCBI BioCollections Database. *Nucleic Acids Research* 51, D1 (2023), D762–D767. https://doi.org/10.1093/nar/gkac1071
[9] Marc Fiume, Miroslav Cupak, Stephen Keenan, Jordi Rambla, Sara de la Torre, Somesh Dyke, Anthony J. Brookes, Daniel Aziz, Soumitra Nag, James Lawson, and et al. 2019. Federated discovery and sharing of genomic data using Beacons. *Nature Biotechnology* 37, 3 (2019), 220–224. https://doi.org/10.1038/s41587-019-0046-x
[10] Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data* 24, 1 (2023), 25.

[11] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. 2009. The Automation of Science. *Science* 324, 5923 (2009), 85–89.

[12] Rodney Alan Long, Shannon Ballard, Syed Shah, Owen Bianchi, Lietsel Jones, Mathew J. Koretsky, Nicole Kuznetsov, Hirotaka Iwaki, and et al. 2024. A new AI-assisted data standard accelerates interoperability in biomedical research. *medRxiv* (2024). https://doi.org/10.1101/2024.10.17.24315618 Preprint (posted Nov 7, 2024).

[13] Patrick Lopes, Uma Radhakrishnan, Peter Alexandrov, M Arif Arshad, Kathryn Beal, Andrew Boland, Martin Brown, Tony Burdett, Claudia Cava, Fiona Cunningham, Dimos Danis, Eimear Dunlea, Philip Ewels, Anne-Gaelle Fenech, Antoine Garcia, Carlos Garcia Giron, Carlos Giron, Syed Haider, Indrani Halder, Chris Hardy, Jennifer Harrow, Angela Hesketh, Graham Hoad, Claire Hunter, Shilpa Iyer, Rohan Jayathilaka, William Jenkinson, Raquel Jimenez, Abdul Karim, Saba Khader, Nikolai Kolesnikov, Gautier Koscielny, Natalia Kurbatova, Kevin Lalle, James Lee, Sha Li, Javier Lopez, Charlotte Lucchetti-Miganeh, Andrew Mallett, Michael Martin, William Mclaren, Kevin Mclay, Andrew Mellor, Jayashree Mistry, Rakesh Nag, Georgia Ntala, Claire O'Donovan, Anika Oellrich, Pall I Olason, Emma Osborne, Paramjit Palasingam, George P Patrinos, George Peat, Gayathri Perera, Michele Pignatelli, Saroj Prachand, Andrew Radford, Marcello Raspa, Laura Rodriguez, Marco Ruffolo, Robert Ryan, Eduardo Sanchez-Garcia, Matthieu Schapira, Paul Schofield, Ana Seabra, Ilya Shmulevich, Varun Singh, James Smith, Richard Smith, Thomas Sneddon, Liliana Sousa, David Spalding, William Spooner, Jen Tang, Aleksandra Tarkowska, Anja Thormann, Sophie To, James Torrance, David Treloar, Chris Turnbull, Meropi Tzouvara, K Joeri van der Velde, Ashish Varma, Shivangi Varma, Suneet Varma, Mark R Viant, Mark Vowles, Linda Wadi, Jasmijn Walter, Jun Wang, Shicong Wang, Ingo Wohlers, David Wratten, Vikas Yadav, Daniel Zerbino, Ahmed Zia, Andrey Zorin, Paul Flicek, Maria Keays, Jyoti Khadake, Julie A McMurry, Phillip Robles, Peter N Robinson, Susanna-Assunta Sansone, Paul N Schofield, Colin L Smith, Nicole L Washington, Monte Westerfield, Ewan Birney, Helen Parkinson, Maria Krestyaninova, Julie A McMurry, Phillip Robles, Peter N Robinson, Susanna-Assunta Sansone, Paul N Schofield, Colin L Smith, Nicole L Washington, Monte Westerfield, Ewan Birney, and Helen Parkinson. 2021. The European Genome-phenome Archive in 2021. *Nucleic Acids Research* 49, D1 (2021), D1023–D1030. https://doi.org/10.1093/nar/gkab1059

[14] Nathan C. Sheffield, Nathan J. LeRoy, and Oleksandr Khoroshevskyi. 2023. Challenges to sharing sample metadata in computational genomics. *Frontiers in Genetics* 14 (2023), 1154198. https://doi.org/10.3389/fgene.2023.1154198

[15] Teknium, Roger Jin, Chen Guang, Jai Suphavadeeprasit, and Jeffrey Quesnelle. 2025. DeepHermes 3 Preview.

[16] Thomas Wolf, Lysandre Debarre, Victor Sanh, Julien Chaumond, Cl'ement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45. https://aclanthology.org/2020.emnlp-demos.6