Bioinformatics, 2024, **40(10)**, btae500
https://doi.org/10.1093/bioinformatics/btae500
Advance Access Publication Date: 8 August 2024
Applications Note

Data and text mining

GENEVIC: GENetic data Exploration and Visualization via Intelligent interactive Console

Anindita Nath [1], Savannah Mwesigwa[1], Yulin Dai [1], Xiaoqian Jiang [2], Zhongming Zhao [1,3,*]

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States
[2]Department of Health Data Science and Artificial Intelligence, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, United States
[3]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, United States

*Corresponding author. Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, United States. E-mail: zhongming.zhao@uth.tmc.edu

Associate Editor: Jonathan Wren

Abstract

Summary: The vast generation of genetic data poses a significant challenge in efficiently uncovering valuable knowledge. Introducing GENEVIC, an AI-driven chat framework that tackles this challenge by bridging the gap between genetic data generation and biomedical knowledge discovery. Leveraging generative AI, notably ChatGPT, it serves as a biologist's "copilot." It automates the analysis, retrieval, and visualization of customized domain-specific genetic information, and integrates functionalities to generate protein interaction networks, enrich gene sets, and search scientific literature from PubMed, Google Scholar, and arXiv, making it a comprehensive tool for biomedical research. In its pilot phase, GENEVIC is assessed using a curated database that ranks genetic variants associated with Alzheimer's disease, schizophrenia, and cognition, based on their effect weights from the Polygenic Score (PGS) Catalog, thus enabling researchers to prioritize genetic variants in complex diseases. GENEVIC's operation is user-friendly, accessible without any specialized training, secured by Azure OpenAI's HIPAA-compliant infrastructure, and evaluated for its efficacy through real-time query testing. As a prototype, GENEVIC is set to advance genetic research, enabling informed biomedical decisions.

Availability and implementation: GENEVIC is publicly accessible at https://genevicanath2024.streamlit.app. The underlying code is open-source and available via GitHub at https://github.com/bsml320/GENEVIC.git (also at https://github.com/anath2110/GENEVIC.git).

1 Introduction

Generative AI, notably Chat GPT and GPT-4 (Achiam *et al.* 2023), excels in a wide array of natural language processing tasks, radically altering knowledge access and processing. Despite the surge in biomedical knowledge and tools, leveraging these advancements fully often requires deep domain and data science expertise, a challenge for many researchers. In addition, these AI models, while capable of generating seemingly accurate outputs, can sometimes produce fictitious information, known as the "hallucination effect" (Rohrbach *et al.* 2018, Xiao and Wang 2021), complicating their reliability in specialized fields. To address this, integrating large language models with domain-specific databases or websites can enhance response accuracy in areas such as genomics (Jin *et al.* 2024), health management (Issom *et al.* 2021), and literacy (Mokmin and Ibrahim 2021).

Here, we introduce GENEVIC, an intelligent interactive console powered by Azure Open AI's generative AI for genetic data exploration and visualization. To demonstrate its utility, we focus on user engagement with the information from Polygenic Score (PGS) Catalog (Lambert *et al.* 2021), a web-based database of polygenic risk scores (https://www.pgscatalog.org). This database is driven by users' deposition of published PGS values that include the variants, alleles, and weights in various phenotypes (e.g. complex diseases), providing a natural resource for GENEVIC to extract prioritized disease-relevant variants. Accordingly, GENEVIC aids in extracting pivotal genetic variants related to diseases, paving the way for building an extensive map of variant-gene-trait associations.

PGS quantifies an individual's genetic predispo sition to a specific disease or trait by aggregating genome-wide genotype effects, derived from genome-wide association studies (GWAS) data (Choi *et al.* 2020). While PGS has been shown promising to evaluate disease risk recently (Lambert *et al.* 2021), highlighted the lack of standardized reporting practices, which hampers PGS research progress. The PGS Catalog addresses this by providing a compre hensive repository of published PGS values, complete with essential metadata for accurate application and assessment, fostering reproducibility. Yet, the PGS Catalog's static format limits interactive exploration, impeding the in-depth analysis of its rich dataset. Importantly, it lacks the capability for data consolidation,

Received: 29 March 2024; **Revised:** 25 June 2024; **Editorial Decision:** 26 July 2024; **Accepted:** 15 August 2024

preventing the integration of various analyses such as network analyses, literature searches, and vi sualization for the related traits. This gap underscores the need for interactive tools like GENEVIC that enable dynamic engagement with the PGS Cat alog, thereby enhancing data mining and the discernment of intricate patterns. GENEVIC not only facilitates detailed analysis and informed decision-making for researchers but also democratizes access to genetic research insights, allowing educators and laypersons to delve into variant functions and hypothesize to advancing the exploration of genetic knowledge.

GENEVIC enhances AI by integrating bioinformatics APIs like STRING and ENRICHR, and supports literature searches from major sites, enriching prioritized variants with specialized knowledge. This integration streamlines research, aiding in the identification of genetic markers and pathways. This a novel prototype of an end-to-end solution that incor porates knowledge, data, and agents into one single portal for a specific domain and thus, broadening access to domain expertise, simplifying data for diverse research backgrounds and promoting multidisciplinary collaboration.

2 Overview of GENEVIC

2.1 Functionality

GENEVIC is a user-friendly intelligent interactive console (or interface) implemented with computational methods to facilitate genetic research.

At its core, the PGSChat component allows researchers to input specific genetic data points, such as single nucleotide polymorphisms (SNPs, including rsIDs and genomic coordinates), gene symbols, and disease or phenotype names as part of prompts, to retrieve relevant information from the PGS rank database, which is a SQLite database that houses variant rankings derived from the PGS Catalog. This interface connects to a GeneAPI Chat interface, utilizing Enrichr and STRING web APIs for enrichment analysis using popular gene set libraries, and gener ating gene-gene interaction networks as well as visualizing the corresponding network graphs, respectively, for a set of genes provided as input. Concurrently, the Literature Search component streamlines the search and retrieval of scientific literature from PubMed, Google Scholar, and arXiv, for a given search query/prompt. In addition, this component can retrieve the abstracts of an article for a given link to that article.

2.2 Implementation

GENEVIC's architecture (Fig. 1) is split between a responsive front end, facilitating user interaction and visualization, and a robust back end, where data processing and API integration occur by harnessing and integrating the power of generative AI. Test cases showcasing sample prompts and corresponding GENEVIC's outputs for each of the three task com ponents are shown in Fig. 1I–III, respectively.

The frontend/user interface was developed using Streamlit. The backend technology of our application is anchored by Azure OpenAI's ChatGPT 3.5 (or GPT-4, if available), a state-of-the-art generative AI model.

Generative AI is pivotal in this application, utilizing prompt-based few-shot learning to classify user prompts for smooth navigation and function execution. It efficiently translates English prompts into SQL and Python codes for querying the PGS rank database and creating intuitive data visualizations. This technology effectively connects to web APIs and enrichment tools with minimal prompts and simplifies literature searches across websites, demonstrating ChatGPT's versatility in managing various tasks and queries.

2.3 PGS rank database

Our study harnessed PGS files from PGS Catalog (Release: 4 August 2023) harmonized to human reference genome (GRCh38 build), creating a local database for easier access. We extracted trait-specific data using R's "Quincunx" package version 1.1.14 (https://github.com/tidyverse/dplyr) (Magno *et al.* 2021), querying the catalog's REST API with the mapped traits (ontology) as listed in Table 1. To ensure consistency amid the dataset's diversity, we focused on essential columns: effect allele, effect weight, and SNP rsID/SNP coordinates, and harmonized variant labeling discrepancies. We merged phenotype-specific PGS files into a single dataset using R's "dplyr" (Wickham *et al.* 2023) and "bind_rows" methods and treated missing values as "NA."

We used the Dowdall method (Fraenkel and Grofman 2014), an alternative Borda method, for rank aggregation of the variants across multiple PGSs based on the absolute values of their effect weights, and annotated the variants using ANNOVAR (Wang *et al.* 2010). The rank aggregation step involved as signing the reciprocal of ranks (RR) to each variant. This means that the top rank receives a RR of 1, the second rank receives a RR of 1/2, the third rank receives a RR of 1/3, and so forth. For each variant query, we calculate the mean of the reciprocals of its ranks (MRR) across multiple PGSs. In cases where weights are unavailable, we assign an MRR of 0. The aggregated ranking of variants by effect weight thus assigns a higher score to variants that have a higher effect weight consistently across multiple PGSs and streamlines querying of trait-relevant top-ranked variants, offering a means to assess the functionality of the PGS Chat feature.

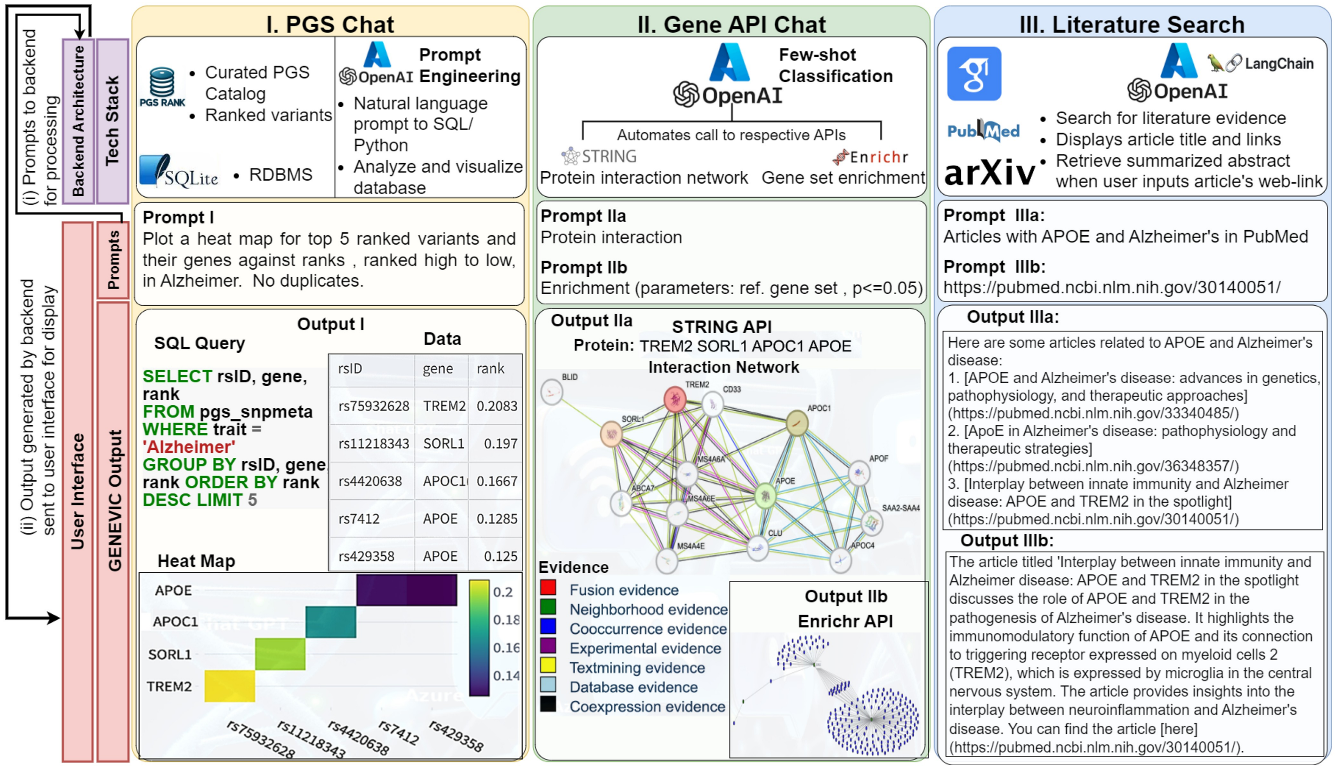
This database was developed with SQLite for its efficiency. Users can switch to any custom database. Currently, GENEVIC supports SQLite and SQL Server, but it can be expanded to other databases.

Download our PGS rank database files and design schema from: https://tinyurl.com/PGSrankDatabase.

2.4 Installation

GENEVIC is accessible for free on the Streamlit community cloud at https://genevicanath2024.streamlit.app, ready to use without any installation. Due to Streamlit's 1 GB data limit, the PGS rank database is restricted to the top 100 genes for Alzheimer's disease, schizophrenia, and cognition, totaling top 300 genes, which is ade quate for evaluating this pilot project. However, users can bypass this limitation by running GENEVIC locally, allowing the upload and analysis of the original comprehensive version of the PGS database that has all the genes for each trait. Taking advantage of this local installation, users can upload and analyze any custom database of any size (currently tested for databases of size over 1.5 GB). Detailed instructions for local installation are at https://tinyurl.com/LocalInstallGuide that has only Python 3.10 or higher as a pre-requisite. In addition, a docker image of GENEVIC is available at https://tinyurl.com/DockerImgIns, with installation and usage instructions. This flexibility also ensures effective use beyond the limitations of the Streamlit community cloud.

Users must have an active Azure Open AI subscription (https://tinyurl.com/AzureOAInst), deploy a ChatGPT or



Prompt I
Plot a heat map for top 5 ranked variants and their genes against ranks , ranked high to low, in Alzheimer. No duplicates.

SQL Query

```
SELECT rsID, gene,
rank
FROM pgs_snpmeta
WHERE trait =
'Alzheimer'
GROUP BY rsID, gene,
rank ORDER BY rank
DESC LIMIT 5
```

Data

rsID	gene	rank
rs75932628	TREM2	0.2083
rs11218343	SORL1	0.197
rs4420638	APOC1(	0.1667
rs7412	APOE	0.1285
rs429358	APOE	0.125

Prompt IIa
Protein interaction

Prompt IIb
Enrichment (parameters: ref. gene set , p<=0.05)

Prompt IIIa:
Articles with APOE and Alzheimer's in PubMed

Prompt IIIb:
https://pubmed.ncbi.nlm.nih.gov/30140051/

Output IIIa:

Here are some articles related to APOE and Alzheimer's disease:
1. [APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches] (https://pubmed.ncbi.nlm.nih.gov/33340485/)
2. [ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies] (https://pubmed.ncbi.nlm.nih.gov/36348357/)
3. [Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight] (https://pubmed.ncbi.nlm.nih.gov/30140051/)

Output IIIb:

The article titled 'Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight discusses the role of APOE and TREM2 in the pathogenesis of Alzheimer's disease. It highlights the immunomodulatory function of APOE and its connection to triggering receptor expressed on myeloid cells 2 (TREM2), which is expressed by microglia in the central nervous system. The article provides insights into the interplay between neuroinflammation and Alzheimer's disease. You can find the article [here] (https://pubmed.ncbi.nlm.nih.gov/30140051/).

Figure 1. Workflow overview between the front-end user interface and the back-end architecture of GENEVIC: (i) user prompts from the interface are sent to the back end for processing and then, (ii) generated output is sent back to the user. The backend comprises common AI services, supported by data tools exclusive to each of the three functionalities: PGS Chat (I), GENEAPI Chat (II), and Literature Search (III). The user prompts: Prompt Ia, Prompt IIa, Prompt IIb, Prompt IIIa and Prompt IIIb generate GENEVIC outputs: Output Ia, Output IIa, Output IIb, Output IIIa and output IIIb, respectively

Table 1. Summary of statistics of the phenotypes used to develop the PGS rank database.[a]

Phenotype (mapped trait)	# PGS files	SNP info
Alzheimer's disease	23	w: [(−0.95)–1.64]; m: 0.0013; md: 5.82e−7; sd: 0.0390
Schizophrenia	5	w: [(−0.04)–0.05]; m: 5.61e−5; md: 8.36e−6; sd: 0.0018
Cognition	5	w: [(−3.60)–1.68]; m: (−0.0005); md: −7e−8; sd: 0.0321

[a] Phenotype denotes disease or trait name. #PGS files denote the count of PGS files for each PGS ID corresponding to the phenotype. SNP info refers to the aggregated information regarding the SNPs or variants in each PGS file. SNP: single nucleotide polymorphism; w: weight range; m, md, and sd denote the mean, median, and standard deviation of the weights, respectively.

GPT-4 model, and enter their account details in GENEVIC's "Settings."

2.5 Test cases

GENEVIC's efficiency was evaluated through simulated real-world research tasks. A comprehensive video walkthrough on utilizing GENEVIC is available at https://tinyurl.com/VideoNavigate. We used ChatGPT 3.5–16k model for our test purposes. GENEVIC leverages Domain-Specific Retrieval Augmented Generation (RAG) to enhance factual accuracy by integrating LLMs with curated databases, external sources such as bioinformatics APIs, and literature sites, ensuring responses are based on verified information.

Documented at https://tinyurl.com/Test-UseCases, each test case confirmed GENEVIC's capability to enhance research efficiency, offering rapid data access and effective visualization, vital in fast-paced research environments.

While GENEVIC may occasionally struggle with precise result retrieval from the PGS rank database due to vague prompts, such challenges can be overcome by refining the prompts, enabled by GenAI's inherent iterative in-context learning capability. For example, enhancing the specificity of the prompt from "Show me the top 10 ranked genes in Alzheimer, top to bottom," to "If duplicate, show once," not only refines the search but also significantly improves the accuracy and relevance of the results, demonstrating GENEVIC's adaptability in navigating complex datasets.

We propose a reliability score based on factual correctness. This involves manually validating GENEVIC's outputs by (i) cross-referencing results from database or Python IDEs using GENEVIC's auto-generated codes, and (ii) comparing API outputs (e.g. STRING, ENRICHR) and literature search results with their original sources. Preliminary evaluation with 10 test cases showed 96% correctness for PGS Chat (using prompt-refinement), 100% for GeneAPI Chat, and 99% for Literature Search.

3 Discussion

3.1 Principle contributions

GENEVIC highlights the capability of cutting-edge generative AI to unify and streamline access to, navigation of and automate analysis of biomedical databases and external web APIs, marking a significant advancement. This platform with an intuitive and user-friendly interface, operates seamlessly, without requiring users to have specific technical or biomedical knowledge or training. It leverages elements of "RAG" via integration of standardized curated databases or real-time information from reliable APIs or sites to ensure accuracy and minimize the risk of AI-generated misinformation, or "hallucination." In addition, GENEVIC allows users to customize data sources and ensures data security through Azure OpenAI's HIPAA-compliant infrastructure, protecting sensitive clinical data.

3.2 Limitations and future directions

This pilot framework, GENEVIC, is only in its nascent stage, designed to evolve with advancements in ChatGPT and related technologies. Currently, it demonstrates capabilities using a limited PGS rank database with data for only three phenotypes and a basic approach to ranking variants via PGS effect weights, highlighting the need for comprehensive data, robust weighing scheme and consideration of various factors such as ethnic background, genotype data, specific PGS scoring techniques, and the degree of sample overlap across PGS datasets for each phenotype, among others.

Future enhancements will broaden the database scope, integrate additional biomedical web APIs into GeneAPI Chat, and enhance Literature Search functionalities by enabling auto-extraction of deeper insights. We plan to introduce automated predictive modeling using generative AI promise to significantly boost GENEVIC's functionality. In future versions, more extensive and external evaluations need to be incorporated such as Hughes Hallucination Evaluation Model (https://tinyurl.com/metricfuture), adapted to classify the factual consistency of responses as a quantifiable measure of reliability; and feedback from a broader and external user base to continuously improve the system.

Thus, this innovative tool not only streamlines research workflows but also sets the stage for equipping future researchers with sophisticated, data-driven tools in genomics and biomedical research.

Acknowledgements

We thank UT Health Science Center at Houston's technical team for their Azure Open AI setup support and colleagues from the Bioinformatics and Systems Medicine Laboratory and the Department of Health Data Science and Artificial Intelligence for their insightful feedback on this project.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

Z.Z. and X.J. were partially supported by the National Institutes of Health (NIH) grant [U01AG079847]. Z.Z. was also partially supported by NIH grants [R01LM012806, R01LM012806-07S1, and R01DE030122]. We thank the technical support from the Cancer Genomics Core funded by the Cancer Prevention and Research Institute of Texas (CPRIT) [RP180734 and RP240610]. X.J. is a CPRIT Scholar in Cancer Research [RR180012], and was also partially supported by Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, and the NIH grants [R01AG066749, R01AG082721, R01AG066749, U24LM013755, U01TR002062, U01CA274576, and U54HG012510]. The sponsors had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

Data availability

Source codes (https://github.com/bsml320/GENEVIC.git) and supplementary materials at https://github.com/bsml320/GENEVIC_Supplementary.git and at *Bioinformatics* online are publicly available.

References

Achiam J, Adler S, Agarwal S *et al.* GPT-4 Technical Report. 2023. https://doi.org/10.48550/arXiv.2303.08774

Choi SW, Mak TS-H, O'Reilly PF. A guide to performing polygenic risk score analyses. *Nat Protoc* 2020;**15**:2759–72. https://doi.org/10.1038/S41596-020-0353-1

Fraenkel J, Grofman B. Strategic voting and coalitions. *Public Choice* 2014;**28**:1–15. https://doi.org/10.1007/BF01718454

Issom DZ, Hardy-Dessources MD, Romana M *et al.* Toward a conversational agent to support the self-management of adults and young adults with sickle cell disease: Usability and usefulness study. *Front Digit Health* 2021;**3**:600333. https://doi.org/10.3389/FDGTH.2021.600333

Jin Q, Yang Y, Chen Q *et al.* GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* 2024;**40**:btae075. https://doi.org/10.1093/bioinformatics/btae075

Lambert SA, Gil L, Jupp S *et al.* The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* 2021;**53**:420–5. https://doi.org/10.1038/s41588-021-00783-5

Magno R, Duarte I, Maia AT. Quincunx: an R package to query, download and wrangle PGS catalog data. *Bioinformatics* 2021;**38**:294–6. https://doi.org/10.1093/BIOINFORMATICS/BTAB522

Mokmin NAM, Ibrahim NA. The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Educ Inf Technol (Dordr)* 2021;**26**:6033–49. https://doi.org/10.1007/S10639-021-10542-Y

Rohrbach A, Hendricks LA, Burns K *et al.* Object hallucination in image captioning. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, 2018, 4035–45. https://doi.org/10.18653/v1/d18-1437

Wang K, Li M, Hakonarson H. An- NOVAR: functional annotation of genetic variants from next-generation sequencing data nucleic acids research. *Nucleic Acids Res* 2010;**38**:e164. https://doi.org/10.1007/BF01718454

Wickham H, François R, Henry L *et al. dplyr: A Grammar of Data Manipulation*. R package version 1.1.4. https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org. 2023.

Xiao Y, Wang WY. On hallucination and predictive uncertainty in conditional language generation. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, Online Conference. Association for Computational Linguistics, 2021, 2734–44. https://doi.org/10.18653/v1/2021.eacl-main.236

Bioinformatics, 2024, 40, 1–4
https://doi.org/10.1093/bioinformatics/btae500
Applications Note