
Alvessa: An Agentic Evidence-Grounded Research Assistant for Genomics

Ksenia Sokolova
Princeton Precision Health
Princeton University
Princeton, NJ
sokolova@princeton.edu

Sanketh Vedula
Princeton Precision Health
Princeton University
Princeton, NJ
svedula@princeton.edu

Keerthana Nallamotu
Department of Computer Science
Princeton University
Princeton, NJ
kn6412@princeton.edu

Guillermo Sapiro
Department of Electrical and Computer Engineering
Princeton University
Princeton, NJ
guillermos@princeton.edu

Olga G. Troyanskaya
Princeton Precision Health
Princeton University
Princeton, NJ
ogt@princeton.edu

Abstract

The rapid proliferation of genomic data, large-scale experiments, and biological foundation models presents major opportunities for biological discovery, but also creates significant integration challenges. Researchers often face a landscape of heterogeneous databases with inconsistent formats, a challenge compounded by the difficulty of integrating this static knowledge with dynamic predictions from foundation models. Furthermore, knowledge from publications quickly becomes outdated and disconnected from new evidence. To address these gaps, we present Alvessa, a research assistant that orchestrates modular agents to perform user-intent understanding, context-specific tool calling, reasoning, and evidence-backed summarization. Alvessa integrates a diverse array of genetic databases, specialized foundation models, and bioinformatics tools, and dynamically selects tools needed for a given query. Unlike conventional portals that return data without reasoning, or general-purpose language models whose conclusions may be outdated due to static training data, Alvessa actively retrieves relevant evidence for a given query, reasons over it to synthesize a conclusion, and presents both the answer and its supporting evidence in an interactive interface. For reproducible assessment, we introduce GenomeArena, a suite of curated benchmarks for evaluating the framework’s core components, enabling granular measurement of entity extraction, tool selection, reasoning, and evidence verification. Our results show high performance while highlighting limitations when necessary databases are not yet integrated.

1 Introduction

The accelerating pace of genomic research, fueled by advances in sequencing technologies, large-scale biobanks, and increasingly powerful biological foundation models, has generated vast quantities of data and opened new avenues for discovery. Yet the promise of these resources remains constrained by

a fundamental bottleneck: the lack of intelligent, flexible systems that can combine static knowledge bases with dynamic, context-specific reasoning.

Modern biomedical inquiry often requires querying across dozens of heterogeneous databases, each with its own schema, user interfaces, update cycles, and specialized expertise. Conventional interfaces offer limited support for integrative analysis, forcing researchers to manually stitch together sources and heuristics, write custom scripts, and spend days sifting through disparate portals. General-purpose language models, while capable of fluent summarization, frequently hallucinate or fail to ground answers in verifiable sources.

To meet this need, we introduce *Alvessa*, an agentic research assistant that performs modular reasoning over structured databases, foundation model predictions, and experimental knowledge. *Alvessa* infers user intent, selects and executes the right tools for the task, maintains provenance of all intermediate artifacts, and synthesizes responses that are explicitly supported by the underlying data, which is surfaced to the user in an interactive interface. This modular design makes the system extensible as new tools, datasets, and models emerge, while its evidence-first execution mitigates hallucination and promotes reproducibility.

To systematically assess these capabilities, we develop *GenomeArena*, a curated benchmark suite that probes the core competencies required for trustworthy genomic assistance: precise entity extraction from noisy queries, accurate and parsimonious tool selection, coherent multi-step reasoning across heterogeneous outputs, and rigorous evidence verification. Our evaluations show strong performance across these dimensions consistently surpassing general-purpose large language models (LLMs).

We view *Alvessa* operating as infrastructure within the genomics ecosystem: an agentic, evidence-grounded assistant for modular reasoning across heterogeneous resources, and *GenomeArena* as a continuously maintained evaluation benchmark that surfaces coverage gaps and guides prioritization of future database and tool integrations.

2 Background

Modern genomics operates within an ecosystem of specialized databases that evolve on different schedules, use incompatible schemas and require a lot of specialized knowledge to use. Primary resources curate or generate domain-specific facts, such as gene and protein interactions in BioGRID Oughtred et al. [2021] and pathway and reaction knowledge in Reactome Milacic et al. [2023], while others organize data at scale, for example protein functions in UniProtKB The UniProt Consortium et al. [2024] and single-cell expression data in CellxGene Program et al. [2023]. Separate portals focus on access and presentation: some take structured inputs like gene sets or variant lists to return summaries and visualizations, and others provide integrated, gene- or target-centric views such as GeneCards Stelzer et al. [2016] and the Open Targets Platform Buniello et al. [2025]. Genome browsers like UCSC Perez et al. [2025] play a dual role by hosting substantial datasets and providing high-throughput visualization. Despite the breadth of information, using it effectively is hard, as identifying relevant information and summarizing it can turn even straightforward questions into days of manual reconciliation.

With the rise of general LLMs, it is tempting to ask them to answer genomic questions directly, and biomedical literature tuned models, such as BioGPT Luo et al. [2022], BioMedLM Bolton et al. [2024], Me-Llama Xie et al. [2025] and others, make that prospect feel close at hand. However, LLMs routinely hallucinate, are easily distracted and project unwarranted confidence even after domain adaptation, a limitation documented broadly and in medical QA and, in part, argued to be intrinsic to current LLMs Xu et al. [2024], Vishwanath et al. [2025]. The problem is amplified in genomics’ long-tail of entities: dbSNP alone now catalogs over 1.2 billion reference SNP identifiers (“rsIDs”) Phan et al. [2024], making memorization and reliable disambiguation unrealistic. Fine-tuning can add domain vocabulary and style, but it does not guarantee answers are grounded in the database records researchers can trust.

A growing set of agentic systems aspires to serve as co-scientists, delivering meaningful gains in planning and protocol support, but most remain literature-centric or coding tools rather than true, database-grounded genomic assistants. Google’s AI co-scientist Gottweis et al. [2025] exemplifies this ambition with multi-agent hypothesis generation and expert-in-the-loop validation, while CRISPR-GPT Qu et al. [2025] automates end-to-end gene-editing design and analysis; both showcase planning and tool use but still lean heavily on text corpora and predefined workflows. In genomics-specific

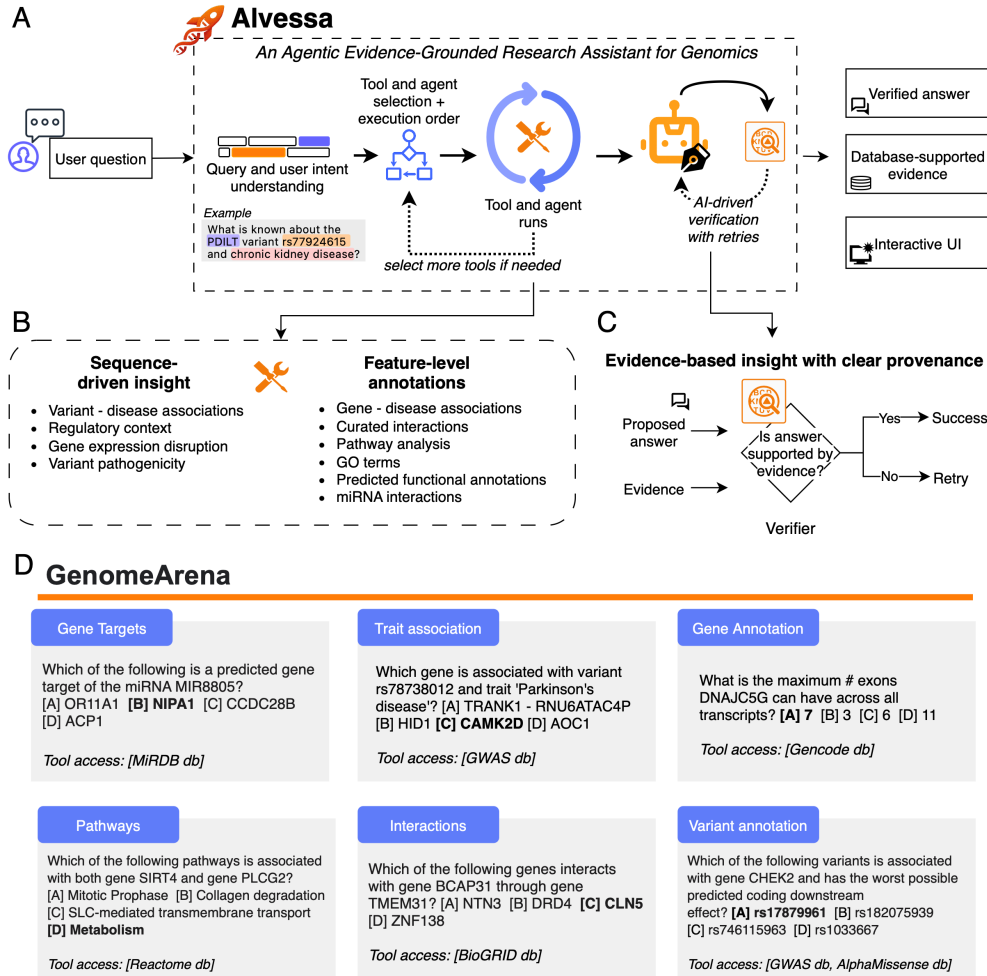


Figure 1: Overview of Alvessa and GenomeArena: (A) System architecture: Alvessa takes user’s natural-language question about variants, diseases or a scientific hypothesis, extracts entities, and selects relevant tools and agents to execute based on user intent. Results are passed through an AI-driven verifier that ensures evidence is grounded in retrieved data; if evidence is insufficient, retries are triggered until a verified answer with clear basis is produced or number of retries is exceeded. (B) Areas that are covered by available agents, spanning sequence-driven features (e.g., variant-disease associations, regulatory context, pathogenicity) and feature-level annotations (e.g., gene annotations, curated interactions, pathway analysis, predicted functional roles). (C) Verification workflow: proposed answers are explicitly checked for database-supported evidence before acceptance. (D) Overview of the GenomeArena, showing diverse tasks (miRNA targets, trait associations, gene annotation, pathway enrichment, interactions, variant annotation) and sample questions. Each question also comes with a record of the database that can be used to answer the question.

settings, GeneAgent Wang et al. [2025] reduces hallucinations by verifying gene-set LLM generated summaries against curated databases, yet its scope is pathway-style analysis and the final outputs do not expose underlying database evidence inline, limiting use outside those niches. Biomni Huang et al. [2025] introduces a general-purpose biomedical agent that plans and executes code over a curated environment of tools and databases, advancing automated research assistance across multiple domains, while Robin Ghareeb et al. [2025] demonstrates multi-agent hypothesis generation with lab-in-the-loop validation toward therapeutic discovery. Our work is complementary: instead of generalist code execution or literature-first discovery, we deliver a genomics-first system whose tools connect to databases directly, with a verifier agent that ties each claim of the response to the returned records and makes them available in the interactive UI for inspection and follow-up.

3 Alvessa’s design

Alvessa is a multi-agent framework designed to perform structured, evidence-grounded reasoning for complex genomics queries. The system follows an iterative workflow of query deconstruction, adaptive tool selection, evidence synthesis, and iterative verification to produce reliable, evidence-grounded answers (Figure 1, Panels A-C). The following sections detail each component of this pipeline.

Query understanding. The initial stage of the Alvessa pipeline deconstructs a user’s natural language query to identify core biological entities. This step is critical as it informs all downstream processes, particularly tool selection. To ensure robust and comprehensive entity recognition, we employ an ensemble of three distinct models: (i) GLiNER Zaratiana et al. [2024], a general-purpose Named Entity Recognition (NER) model; (ii) Flair Weber et al. [2021], a domain-adapted model optimized for biological text; and (iii) a large language model (claude-3-haiku) for broad coverage. In addition, we perform rule-based extraction of structured identifiers, including Ensembl IDs for genes, transcripts, and proteins, and dbSNP rsIDs for genetic variants. The final list of entities is generated from the union of all entities identified by these models. This consolidated set serves as the primary input for the subsequent tool selection agent.

Tool selection. To handle the heterogeneity of biological questions, Alvessa incorporates an adaptive tool-selection process. Based on the entities identified from the user query, a *tool-selection agent* evaluates the user’s query, the extracted entities, and the descriptions of available tools. In addition the model is provided with an example of the question-tool combinations. This initial step determines which external databases or predictive models are most relevant to run first to answer the query. To ensure this tool selection is not static, after the initial tools are executed, the system state and the original question are passed to another reasoning model instance. This model then determines whether additional tools are necessary to answer the query. This approach ensures both efficiency, by avoiding unnecessary calls, and completeness, by expanding the evidence base only when required.

Synthesis. Following tool execution, the collected data, ranging from curated interaction data to predicted regulatory effects, are aggregated into a structured system state. Before synthesis, a condensation step refines this information, trimming verbose outputs and reformatting key data into concise, natural-language summaries. This structured context is then provided to a *synthesis agent*, which integrates the heterogeneous evidence to generate a candidate answer. To enforce factual grounding, the agent’s output is constrained to a structured format containing two key fields: the natural-language answer and a list of the specific data entries that support it. This design allows Alvessa to move beyond simple data retrieval and produce biologically coherent narratives that are explicitly linked to their underlying evidence.

Verification. To ensure the reliability and accuracy of its outputs, every candidate answer undergoes a stringent verification process. A *verification agent* assesses whether the generated answer is fully and explicitly supported by the evidence cited by the synthesis agent (Figure 1, Panel C). It checks for logical consistency, factual accuracy, and absence of hallucination or unsupported claims. If the answer lacks sufficient grounding, it is rejected. The framework then re-invokes the synthesis agent with the directive to generate a new answer based on the same evidence, effectively entering a refinement loop. This process continues until a verifiable, evidence-grounded answer is produced or a predefined retry limit is reached. Only answers that successfully pass this verification stage are returned to the user, guaranteeing that all outputs are trustworthy and evidence-based.

4 Tool catalog

Alvessa integrates a diverse suite of tools to provide comprehensive biological context for genes and variants (Figure 1, Panel B). These tools are not all executed for every query; instead, a dynamic tool selector determines which tools are most appropriate to call based on the specific biological question. The available tools are organized into three main categories: curated databases, predicted functional and regulatory annotations, and genomic foundation models.

4.1 Curated databases

This category includes established databases that provide manually curated and widely accepted information on gene-trait associations, biological pathways, molecular interactions, and genomic structures. These tools form the foundation of our evidence-based annotations and can be called upon to retrieve foundational knowledge.

Genome Wide Association Studies (GWAS). For comprehensive gene-trait associations, Alversa can query the GWAS Catalog Buniello et al. [2018], a publicly available database of published genome-wide association studies. This resource contains over 880k associations, each detailed with the risk score (e.g., odds ratio or beta coefficient), study sample size, and PubMed ID of the publication. We consider associations with a p-value less than the conventional significance threshold of 5×10^{-8} , resulting in a set of 500k significant associations.

Pathway and functional annotations. Pathway-level context can be provided by Reactome, which links genes to their respective biological pathways, deduplicated for a concise representation. Further functional detail may be derived from Gene Ontology (GO) Ashburner et al. [2000] terms queried from UniProt. A GO summarization tool distills these terms to highlight gene’s key biological roles.

Gene interaction networks. Alversa utilizes BioGRID to retrieve curated interaction partners for a gene of interest. The framework queries BioGRID’s API and organizes interaction data by the experimental system used for detection. While primarily focused on human data, selected non-human interactors, such as those from synthetic lethality studies, are also included.

Gene structure and variant annotations. Information on gene structure, including genomic location, number of transcripts, and their complexity (e.g., number of exons), can be retrieved using Gencode Mudge et al. [2025]. To annotate genetic variants, the system leverages the Single Nucleotide Polymorphism Database (dbSNP) Phan et al. [2024], a public archive of short genetic variations. This tool is used to retrieve precise genomic coordinates (chromosome, position, reference/alternative alleles) mapped to specific reference genome assemblies (e.g., GRCh37/GRCh38). Furthermore, it can provide population-specific allele frequencies, offering essential context on variant prevalence across diverse global populations.

4.2 Predicted functional and regulatory annotations

To supplement curated data, Alversa can call upon tools that provide computationally predicted annotations for gene function and regulatory interactions, offering insights beyond experimentally confirmed knowledge.

HumanBase. Alversa is integrated with HumanBase, a resource that provides data-driven predictions of gene function, regulation, and interactions across human tissues and cell types. Specifically, we use it to obtain predicted functional annotations per gene, subsequently filtered to retain only the most confident predictions (≥ 0.9). These represent likely gene functions beyond direct experimental evidence.

miRNA targets. The miRDB tool adds regulatory perspective by annotating miRNA entities with their predicted gene targets using the miRDB v6.0 database Chen and Wang [2019]. By incorporating miRNA-gene regulatory links across species, this tool extends Alversa’s ability to connect noncoding RNA activity with gene-level functional outcomes.

4.3 Genomic foundation models

Alversa also leverages state-of-the-art deep learning models to interpret the functional consequences of genetic variation, providing predictive insights when needed.

AlphaMissense. The pathogenicity of missense variants can be assessed using AlphaMissense Cheng et al. [2023], a deep learning predictor that integrates a protein language model, structural features, and population frequency data. Although variant locations are retrieved for both GRCh37 and GRCh38 assemblies, AlphaMissense operates on GRCh38 coordinates. Since scores are gene-specific, a match of both the coordinate and gene ID to the gene of interest is required. When the tool is called, pathogenicity predictions are added for every SNP in the state that has a valid AlphaMissense score.

Sei. For non-coding variants, Alvessa can employ Sei Chen et al. [2022], a sequence-based foundation model that predicts regulatory activity. Sei classifies any given genomic sequence into one of 40 "sequence classes," allowing the region surrounding a variant to be annotated with its most likely regulatory function.

ExpectoSC. To predict the impact of variants on gene expression, Alvessa can apply ExpectoSC Sokolova et al. [2023], a deep learning model trained on sequence data across 105 cell types. The model scores variants in a 40 kb window around gene’s transcription start site, and its predictions are z-score normalized against a background of common variants. The cell types predicted to have the strongest up- or down-regulation effects are highlighted for further analysis.

5 GenomeArena

To rigorously assess the performance of our agentic system, we introduce GenomeArena, a benchmark structured into two complementary parts: (i) *prerequisite skills* (entity extraction and tool selection) that ensure reliable grounding and orchestration; and (ii) *task-oriented biological benchmarks* that simulate real-world queries (Figure 1, Panel D).

Entity extraction. Correctly resolving biological entities (genes, transcripts, variants, miRNAs, traits/diseases, genomic loci) is a prerequisite for any downstream reasoning. This benchmark evaluates whether the system can robustly identify mentions of aforementioned entities independent of any subsequent biological tool use.

Tool selection. Beyond recognition, effective assistance depends on choosing and sequencing tools appropriately. Given a query and an available toolset, this category measures whether the agent selects a minimally sufficient set, invokes tools in a coherent order, and avoids unnecessary calls, quantifying orchestration skill separately from factual accuracy.

Task-oriented biological benchmarks comprise a multiple-choice question set tailored to genomics. Questions are grouped by the biological task they represent (detailed below) and emphasize realistic, task-driven scenarios that mirror real-world queries. Example questions are presented in Table 1 in the Appendix.

Trait association. Identifying associations between genetic variants and observable traits or diseases is a fundamental task in genetics. This set of questions is designed to evaluate the model’s ability to accurately recover gene-trait associations.

Gene annotation. Understanding gene’s structure, including its various transcripts from alternative splicing, is important for a variety of downstream tasks and experiments. This category focuses on retrieving specific details of gene structure without manual database searches.

Variant effects. Annotating variants with functional impact helps to prioritize them for further study and clinical assessment. This benchmark tests the ability of the system to answer more complex, multi-step questions, streamlining the process of identifying high-priority variants for functional studies and clinical follow-up.

Gene interactions. Proteins function within intricate networks of interactions, and mapping these is essential for understanding cellular processes and molecular mechanisms of disease. Questions in this category probe the system’s ability to analyze biological networks.

Pathway analysis. Contextualizing a gene within its relevant biological pathways is key to understanding its function. This benchmark assesses the ability to place genes in a functional context, giving biologists immediate functional context for a gene of interest, helping to form hypotheses about its role in the cell.

miRNA targets. Identifying gene targets of miRNAs is critical to determine their biological function as well as the role they play in disease onset. Questions in this category test the ability to accurately retrieve and recognize predicted targets for a given miRNA of interest.

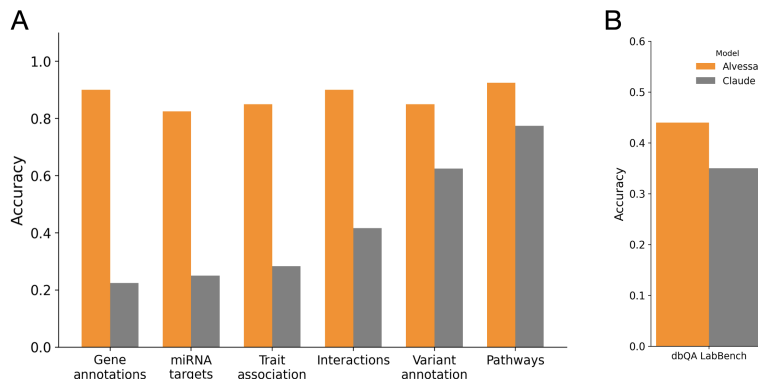


Figure 2: Evaluation of Alvessa against Claude-sonnet-4. (A) Accuracy on the GenomeArena multiple choice questions; (B) Accuracy on the set of 100 dbQA LabBench questions.

6 Results

Across the GenomeArena benchmarks, Alvessa consistently surpasses Claude (Claude-sonnet-4), demonstrating both the reliability of the framework and the value of integrating databases into the reasoning process (Figure 2A).

The largest gains occur in gene annotation and miRNA target prediction. These categories are particularly important because they illustrate cases where language models alone are insufficient. Questions such as the number of exons in a gene or known miRNA targets may appear straightforward, yet LLMs frequently fail to provide accurate answers. Alvessa, by grounding responses in reliable databases, delivers high accuracy in these domains, highlighting the gap between perceived and actual LLM capabilities.

For pathway-level tasks, Claude’s baseline performance is already relatively strong, likely reflecting broader coverage of pathway relationships in training corpora. Even so, Alvessa consistently performs better, with the added advantage of drawing from up-to-date pathway databases rather than relying on static model memorization. This highlights how database integration complements pretrained knowledge by providing current and verifiable information.

Finally, we evaluate Alvessa on a set of 100 dbQA questions from LabBench (Figure 2B). Again, Alvessa clearly outperforms Claude, but the absolute performance is lower than on GenomeArena. This is expected, as some LabBench questions rely on resources not yet included in the system, for example Gene Transcription Regulation Database Kolmykov et al. [2020]. Addressing these gaps will be a critical step in future iterations of the framework.

We next examine tool selection behavior, enabled by the metadata available in GenomeArena. Across tasks, Alvessa’s tool-selection agent is highly reliable, almost always invoking the correct tools (Table A.1 in Appendix). The only systematic deviations occur in the more complex queries requiring both the GWAS Catalog and AlphaMissense; in these edge cases, the model occasionally failed to complete the full sequence of calls. Interestingly, we also observe the inverse pattern: questions that trigger extra, unnecessary tools. This effect is most apparent in pathway queries (Reactome tool), which on average include one additional tool invoked. Closer inspection revealed that alongside Reactome, Alvessa sometimes invoked HumanBase functional annotation tools and UniProt, reflecting partial overlap in the types of evidence relevant to pathway-level reasoning.

A critical component of Alvessa is reliable entity recognition, since downstream tool selection depends on accurate identification of genes, variants, and traits in user queries. To evaluate this step, we benchmarked multiple extraction approaches across eight sets spanning a range of entity types and complexities. Example queries and results are presented in Tables 3 and 4) in the Appendix, respectively. Claude-haiku alone performs competitively on simple cases but shows noticeable drops in accuracy for more complex sets (e.g., set6-set8, multiple genes with inconsistent sentence structure). In contrast, dedicated NER models such as Flair and GLiNER achieve consistently higher accuracy across all sets. The strongest performance is obtained by combining methods: Flair+GLiNER

system can retrieve or reason about. While this is a fundamental limit in data-driven scientific domains, it also points toward important future directions: (i) expanding coverage across additional biological layers (e.g., transcriptional regulation, multi-omics integration), (ii) designing orchestration policies capable of multi-step, hypothesis-level reasoning while remaining data-grounded, and (iii) developing community benchmarks such as GenomeArena into shared standards for evaluating scientific assistants.

Alvessa demonstrates the promise of evidence-grounded, agentic systems in genomics and beyond. By exposing not only answers but also the data behind them, it provides a blueprint for how AI systems can become reliable research assistants: transparent, extensible, data-driven and aligned with the needs of scientific practice.

8 Acknowledgments

This work was supported in part by NIH grant 5U24DK100845, and NIH grant 1U01DK133090 to O.G.T. This research was also supported in part by a grant to support S.V. from the Schwab Charitable Fund made possible by the generosity of Eric and Wendy Schmidt. This work was supported, in whole or in part, by the Gates Foundation [INV-081342]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Additional support comes from NSF, ONR, and the Simons Foundation.

References

- Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi: 10.1038/75556.
- Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D Manning. BioMedLM: A 2.7b parameter language model trained on biomedical text. March 2024.
- Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, and et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1), Nov 2018. doi: 10.1093/nar/gky1120.
- Annalisa Buniello, Daniel Suveges, Carlos Cruz-Castillo, Manuel Bernal Llinares, Helena Cornu, Irene Lopez, Kirill Tsukanov, Juan María Roldán-Romero, Chintan Mehta, Luca Fumis, Graham McNeill, James D Hayhurst, Ricardo Esteban Martinez Osorio, Ehsan Barkhordari, Javier Ferrer, Miguel Carmona, Prashant Uniyal, Maria J Falaguera, Polina Rusina, Ines Smit, Jeremy Schwartzentruber, Tobi Alegbe, Vivien W Ho, Daniel Considine, Xiangyu Ge, Szymon Szyszkowski, Yakov Tsepilov, Maya Ghoussaini, Ian Dunham, David G Hulcoop, Ellen M McDonagh, and David Ochoa. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res*, 53(D1):D1467–D1475, January 2025.
- Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, Jul 2022. doi: 10.1038/s41588-022-01102-2.
- Yuhao Chen and Xiaowei Wang. Mirdb: An online database for prediction of functional microRNA targets. *Nucleic Acids Research*, 48(D1), Aug 2019. doi: 10.1093/nar/gkz757.
- Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, and et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664), Sep 2023. doi: 10.1126/science.adg7492.
- Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. Robin: A multi-agent system for automating scientific discovery. May 2025.

- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist. February 2025.
- Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025. doi: 10.1101/2025.05.30.656746. URL <https://www.biorxiv.org/content/early/2025/06/02/2025.05.30.656746>.
- Semyon Kolmykov, Ivan Yevshin, Mikhail Kulyashov, Ruslan Sharipov, Yury Kondrakhin, Vsevolod J Makeev, Ivan V Kulakovskiy, Alexander Kel, and Fedor Kolpakov. Gtrd: an integrated view of transcription regulation. *Nucleic Acids Research*, 49(D1):D104–D111, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1057. URL <https://doi.org/10.1093/nar/gkaa1057>.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. October 2022.
- Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2024. *Nucleic Acids Res*, 52(D1):D672–D678, November 2023.
- Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, Jose Gonzalez Martinez, Toby Hunt, Irwin Jungreis, Jane E Loveland, Carme Arnan, If Barnes, Ruth Bennett, Andrew Berry, Alexandra Bignell, Daniel Cerdán-Vélez, Kelly Cochran, Lucas T Cortés, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Reham Fatima, Matthew Hardy, Prajna Hebbar, Zoe Hollis, Benjamin T James, Yunzhe Jiang, Rory Johnson, Gazaldeep Kaur, Mike Kay, Riley J Mangan, Miguel Maquedano, Laura Martínez Gómez, Nourhen Mathlouthi, Ryan Merritt, Pengyu Ni, Emilio Palumbo, Tamara Perteghella, Fernando Pozo, Shriya Raj, Cristina Sisu, Emily Steed, Dulika Sumathipala, Marie-Marthe Suner, Barbara Uszczyńska-Ratajczak, Elizabeth Wass, Yucheng T Yang, Dingyao Zhang, Robert D Finn, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L Tress, Ewan Birney, Fergal J Martin, and Adam Frankish. Gencode 2025: reference gene annotation for human and mouse. *Nucleic acids research*, 53(D1):D966–D975, January 2025. ISSN 0305-1048. doi: 10.1093/nar/gkae1078. URL <https://europepmc.org/articles/PMC11701607>.
- Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, Sonam Dolma, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, January 2021.
- G Perez, G P Barber, A Benet-Pages, J Casper, H Clawson, M Diekhans, C Fischer, J N Gonzalez, A S Hinrichs, C M Lee, L R Nassar, B J Raney, M L Speir, M J van Baren, C J Vaske, D Haussler, W J Kent, and M Haeussler. The UCSC genome browser database: 2025 update. *Nucleic acids research*, 53(D1), January 2025.
- Lon Phan, Hua Zhang, Qiang Wang, Ricardo Villamarin, Tim Hefferon, Aravinthan Ramanathan, and Brandi Kattman. The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res*, 53(D1):D925–D931, November 2024.
- CZI Single-Cell Biology Program, Shibli Abdulla, Brian Aeversmann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel

- Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174. URL <https://www.biorxiv.org/content/early/2023/11/02/2023.10.30.563174>.
- Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C Cousins, William A Johnson, Xiaotong Wang, Mihir Shah, Russ B Altman, Denny Zhou, Mengdi Wang, and Le Cong. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, pages 1–14, July 2025.
- Ksenia Sokolova, Chandra L. Theesfeld, Aaron K. Wong, Zijun Zhang, Kara Dolinski, and Olga G. Troyanskaya. Atlas of primary cell-type-specific sequence models of gene expression and variant effects. *Cell Reports Methods*, 3(9):100580, Sep 2023. doi: 10.1016/j.crmeth.2023.100580.
- Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, 54(1):1.30.1–1.30.33, June 2016.
- The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpenter, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu, Rossana Zaru, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echoukh, Elisabeth Coudert, Beatrice Cuhe, Edouard de Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Anastasia Sveshnikova, Cathy H Wu, Cecilia N Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Yuqi Wang, and Jian Zhang. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res*, 53(D1):D609–D617, November 2024.
- Krithik Vishwanath, Anton Alyakin, Daniel Alexander Alber, Jin Vivian Lee, Douglas Kondziolka, and Eric Karl Oermann. Medical large language models are easily distracted. April 2025.
- Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, July 2025.
- Leon Weber, Mario Sanger, Jannes Munchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794, 2021.
- Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(1):1–10, March 2025.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. January 2024.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, 2024.

A Supplemental Materials

A.1 Supplemental Tables

Benchmark	Example query schema	Example question
Protein-protein interactions	Which of the following genes interacts with gene {GENE 1} through gene {GENE 2}? [A] {GENE A} [B] {GENE B} [C] {GENE C} [D] {GENE D}. Answer: {A/B/C/D}	Which of the following genes interacts with gene BCAP31 through gene TMEM31? [A] NTN3 [B] DRD4 [C] CLN5 [D] ZNF138. Answer: C
Trait association	Which gene is associated with variant {VARIANT} and trait {TRAIT}? [A] {GENE A} [B] {GENE B} [C] {GENE C} [D] {GENE D}. Answer: {A/B/C/D}	Which gene is associated with variant rs78738012 and trait <i>Parkinson's disease</i> ? [A] TRANK1-RNU6ATAC4P [B] HID1 [C] CAMK2D [D] AOC1. Answer: C
Gene annotation	What is the maximum # exons {GENE} can have across all transcripts? [A] {COUNT A} [B] {COUNT B} [C] {COUNT C} [D] {COUNT D}. Answer: A	What is the maximum # exons DNAJC5G can have across all transcripts? [A] 7 [B] 3 [C] 6 [D] 11. Answer: A
miRNA targets	Which of the following is a predicted gene target of the miRNA {miRNA}? [A] {miRNA A} [B] {miRNA B} [C] {miRNA C} [D] {miRNA D}. Answer: B	Which of the following is a predicted gene target of the miRNA MIR8805? [A] OR11A1 [B] NIPA1 [C] CCDC28B [D] ACP1. Answer: B
Variant annotation	Which variant is associated with gene {GENE} and has the worst predicted coding downstream effect? [A] {VARIANT A} [B] {VARIANT B} [C] {VARIANT C} [D] {VARIANT D}. Answer: A	Which variant is associated with gene CHEK2 and has the worst predicted coding downstream effect? [A] rs17879961 [B] rs182075939 [C] rs746115963 [D] rs1033667. Answer: A
Pathway analysis	Which pathway is associated with both gene {GENE 1} and gene {GENE 2}? [A] {PATHWAY A} [B] {PATHWAY B} [C] {PATHWAY C} [D] {PATHWAY D}. Answer: D	Which pathway is associated with both SIRT4 and PLCG2? [A] Mitotic Prophase [B] Collagen degradation [C] SLC-mediated transmembrane transport [D] Metabolism. Answer: D

Placeholders in braces (e.g., {GENE}, {VARIANT}) are filled when instantiating a question.

Table 1: Example query templates for the GenomeArena benchmark

Needed tool	Accuracy	Mean # extra tools
BioGRID	1.000	0.0
GWAS	1.000	0.2
Gencode	1.000	0.0
miRNA targets (miRDB)	1.000	0.0
GWAS + AlphaMissense	0.975	0.0
Reactome	1.000	0.8
Overall	0.996	0.16

Table 2: Tool selection performance summary.


Set	Query	Recognized genes
S1	What are variants for gene SLC36A1?	SLC36A1
S2	List common interactions for genes UGT2B11, PAOX, and ACTR8.	UGT2B11, PAOX, ACTR8
S3	List all of the variants in EEF1g.	EEF1g
S4	What is common between NCbp21 and DPM2?	NCbp21, DPM2
S5	PATE3 is important for?	PATE3
S6	What pathways are associated with HCN2 as well as PGF but not with CD300E?	HCN2, PGF, CD300E
S7	Given gene ZNF827, which of the following are interacting? CCR2, DaOA.	ZNF827, CCR2, DaOA
S8	What are the pathways associated with MND1, TBR1, SF1, CWH43, rEG1b, acbd6, KRIT1, OR2T1, nubp1, SPINK1?	MND1, TBR1, SF1, CWH43, rEG1b, acbd6, KRIT1, OR2T1, nubp1, SPINK1

Table 3: Example queries with the ground-truth genes for each set in the entity-extraction benchmark.

Method	S1	S2	S3	S4	S5	S6	S7	S8	Mean
claude-haiku	1.00	0.96	0.96	0.93	0.97	0.79	0.68	0.78	0.88
gliner	1.00	1.00	0.97	0.95	0.99	0.98	0.99	1.00	0.99
flair	0.99	1.00	0.99	0.91	0.95	0.99	0.98	0.97	0.97
claude-haiku + gliner	1.00	1.00	0.99	0.98	0.99	1.00	0.99	0.98	0.99
claude-haiku + flair	1.00	1.00	0.99	0.98	0.98	1.00	0.98	0.97	0.99
flair + gliner	1.00	1.00	1.00	0.97	1.00	0.99	1.00	1.00	1.00
claude-haiku + flair + gliner	1.00	1.00	0.99	0.98	0.99	1.00	1.00	0.99	0.99

Table 4: Accuracy of entity extraction across methods on eight evaluation sets. **Bold** marks the best per column (ties allowed).

B Example UI output



Alvessa: An Agentic Evidence-Grounded Research Assistant for Genomics

Question

Which variants are implicated in IRF5 and cancer and what are their functional consequences?

Detected Entities

Genes: IRF5

Tools Used

alphamissense extract_entities humanbase_expecto humanbase_tissue_expecto_annotate_variants query_gwas_extensive reactome sei uniprot_base variant_annotations BioGRID Summarize_bioGRID_GO alphamissense extract_entities humanbase_expecto humanbase_tissue_expecto_annotate_variants query_gwas_extensive reactome sei uniprot_base variant_annotations variant_population_summaries

Answer

Based on the GWAS data, two IRF5 variants are specifically associated with cancer: rs3807306 and rs2070197. rs3807306 is associated with kidney cancer ($p=8e-18$, risk score=0.91) and clear cell renal cell carcinoma ($p=1e-13$, risk score=0.91). rs2070197 is associated with systemic lupus erythematosus ($p=6e-24$, risk score=1.88). The functional consequences include: rs3807306 is an intronic variant with B-cell-like regulatory activity that shows tissue-specific expression effects, with strongest negative effects in kidney immune cells (MNPb non-classical monocyte derived and B cells). rs2070197 is a 3' UTR variant with monocyte/macrophage regulatory activity, located in a functionally important region that could affect mRNA stability or translation. Both variants show population frequency variations and are predicted to have regulatory roles in immune cell function.

Evidence

rs3807306 associated with kidney cancer ($p=8e-18$, risk score=0.91) and clear cell renal cell carcinoma ($p=1e-13$, risk score=0.91)
rs2070197 associated with systemic lupus erythematosus ($p=6e-24$, risk score=1.88)
rs3807306 is an intron_variant with E5 B-cell-like regulatory activity
rs2070197 is a 3_prime_UTR_variant with E7 Monocyte/Macrophage regulatory activity
rs3807306 shows strongest negative expression effects in MNPb non-classical monocyte derived/kidney (-7.6332) and gd T cell 1/liver (-7.2115)
rs2070197 shows strongest positive expression effects in DC activated/lung (0.2262) and DC 2/lung (0.2254)
IRF5 function: Transcription factor that plays a critical role in innate immunity by activating expression of type I interferon

BioGRID (genes)

IRF5
Reconstituted Complex i
HDAC1 SIN3A CREBBP EP300 IRF3 TRIM21 MAVS ALPI NCOR2 COPS3
Two-hybrid i
SGTA TRIM15
Affinity Capture-Western i
COP55 COP54 MGEA5 ALPI SLC15A4 COP57A EP300 NCOR2 CHUK COP56 NCOR1 GPS1 TRIM35 IRF5 TRAF6 TRIM21
OGT COP53 PELI1 HDAC1 MYD88 SIN3A COP58 BIRC3 COP52 CREBBP SGTA KAT2B
Affinity Capture-MS i
RG53 GPS1 IRF6 COP54 PPM1G PSMB7 CIT GTPBP1 FLOT2 SLC15A4 COP57A GRB2 MAVS ALPI CEP152 GTPBP2 HIVEP1
FRET i
MYD88
Co-crystal Structure i
IRF5
Proximity Label-MS i
PPIB PACSIN2 PPP6R1 PPP6R3 FAM21A ANKRD28 CEP152
Affinity Capture-RNA i
TOP3B

Selected non-human interactions
IRF5
none

Figure 4: Output window for the question: "Which variants are implicated in IRF5 and cancer and what are their functional consequences?"

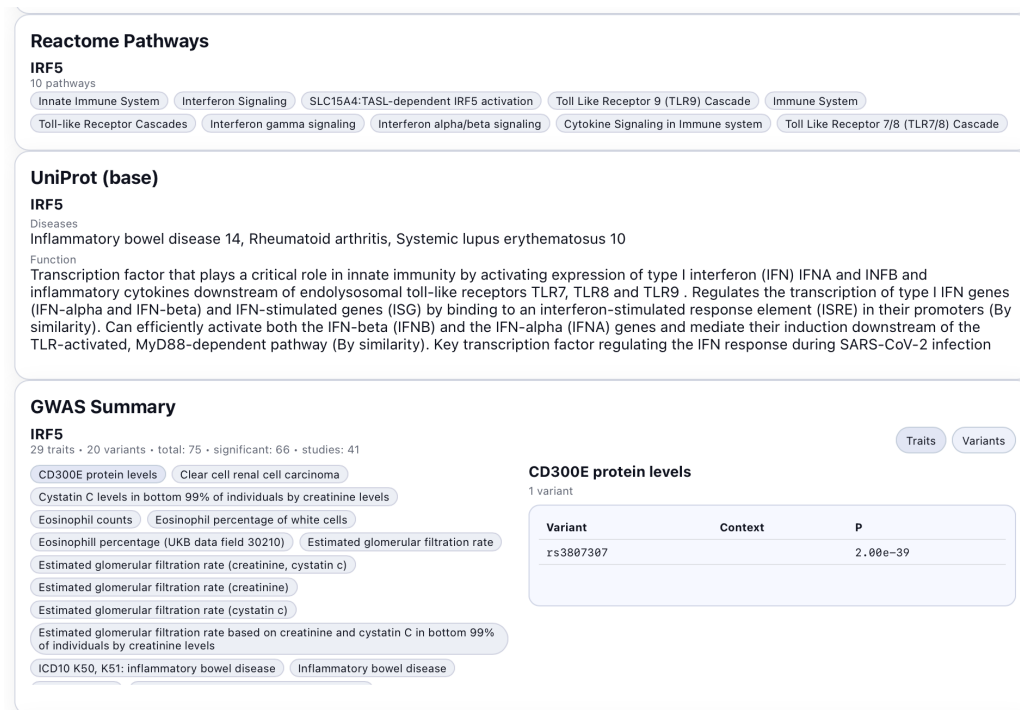


Figure 5: Following from Figure 4: data returned by the tools is presented as evidence together with the synthesized summary.

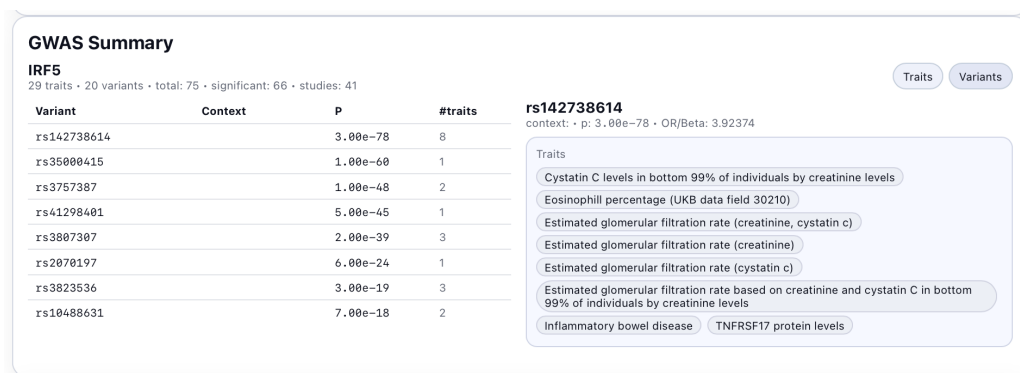



Figure 6: A variant-focused view of the GWAS summary, alternative to the trait-centric view in Figure 5.

rs2070197	other	IRF5	E7 Monocyte / Macrophage	highest 0.226 (78th pct) — DC activated / lung lowest -0.023 (35th pct) — GABAergic interneurons 1 / brain Details	T→C	Syst
rs3823536	intronic	IRF5	E5 B-cell-like	highest 0.759 (91st pct) — MNPb non-classical monocyte derived / kidney lowest -0.131 (9th pct) — Neuronal Stem cells / brain Details	G→A	Eosin SLAM
rs10488631	intergenic	IRF5 - TNPO3	E5 B-cell-like	highest 0.070 (83rd pct) — Pyramidal neurons (hip CA) 3 / brain lowest -0.493 (16th pct) — DC activated / lung Details	T→C	Syst
rs3807306	intronic	IRF5	E5 B-cell-like	highest 0.617 (98th pct) — GABAergic interneurons 1 / brain lowest -7.633 (2nd pct) — MNPb non-classical monocyte derived / kidney Details	G→A G→C G→T	ICD1 Kidn Rheu Rheu
<p>ExpectoSC (cell-type specific expression disruption)</p> <p>chr7:128580680 T</p> <p>n_tissues: 105 • mean: -2.158 • median: -2.044 • min: -7.633 • max: 0.617 • std: 1.992</p> <p>relative effect: lower than average</p> <p>highest 0.617 (98th pct) — GABAergic interneurons 1 / brain</p> <p>lowest -7.633 (2nd pct) — MNPb non-classical monocyte derived / kidney</p> <p>Top tissues</p> <ul style="list-style-type: none"> GABAergic interneurons 1 / brain — score 0.617, 98th pct GABAergic interneurons 2 / brain — score 0.534, 98th pct <p>Lowest tissues</p> <ul style="list-style-type: none"> MNPb non-classical monocyte derived / kidney — score -7.633, 2nd pct gd T cell 1 / liver — score -7.212, 2nd pct 						
rs10954214	other	IRF5	TN1 Transcription	highest 0.093 (82nd pct) — Epithelial progenitor cell / kidney lowest -0.094 (34th pct) — MNPb non-classical monocyte derived / kidney Details	C→T	Syst scler arthr

Figure 7: Example variant summary entry with expanded information on cell type-specific gene expression predictions for rs3807306, illustrating the information context. This variant is associated with kidney cancer and clear cell renal cell carcinoma. Based on the predictions from deep learning sequence models, it is located in the E5 B-cell-like sequence region and its strongest predicted effect is in kidney immune cells (MNPb, non-classical monocyte-derived), where it exhibits severe downregulation. Given its strong regulatory predictions, consistent across multiple foundation models and supported by GWAS disease signals, this variant is of potential interest for follow-up studies. Alvessa highlighted this variant in its answer as one of potential interest.



Alvessa: An Agentic Evidence-Grounded Research Assistant for Genomics

Enter a question to explore entities, tools, and curated outputs.

Which variants are implicated in IRF5 and cancer and what are their functional consequences?

Run

Tip: press Enter to run

Figure 8: Alvessa's landing page.

B.1 Evaluation details

To convert the model into multiple-choice system for GenomeArena and LabBench, the following system prompt was used:

```
You are a multiple-choice answering system. You must reply
with exactly one of the following letters: A, B, C, or D. Do
not include any explanation, reasoning, or extra text. Your
response will be parsed by a program that will fail if you
output anything other than a single capital letter. Example
valid output: C Example invalid outputs: "Answer: C", "C.",
"Option C", "B because...", "A is correct because..."
```

For the benchmark tasks, we simplified the Alvessa pipeline by disabling both the secondary tool-selection step (beyond the first iteration) and the verifier agent. Each prompt was presented to the system in multiple-choice text format, with candidate answers labeled A, B, C, D, as shown in examples. Both Claude and Alvessa models occasionally returned free-text responses despite the prompt structure, and this we considered these valid if the final word of the output was one of the four options.

B.2 Example query

Additional sample query with response and evidence, shown as plain text

Sample Output: RICTOR

Question

Through which mechanisms and variant effects is RICTOR connected to cancer and metabolite measurements?

Detected Entities

Genes: RICTOR

Tools Used

BioGRID, Summarize BioGRID GO, Extract Entities, Gencode, HumanBase Functions, Query GWAS by Gene, Reactome, UniProt Base, UniProt GWAS, ExpectoSC, Variant Annotations, Variant Population Summaries

Answer

RICTOR is connected to cancer and metabolite measurements through multiple mechanisms involving mTORC2 signaling, AKT pathway regulation, and metabolic control. The primary mechanisms include: (1) mTORC2 complex function where RICTOR acts as a molecular adapter in the mechanistic target of rapamycin complex 2, which transduces growth factor signals to proliferation and anabolic pathways; (2) AKT kinase activation through phosphorylation of AKT1 at multiple sites including Ser-473, facilitating full AKT activation for cancer-related signaling; (3) metabolic regulation via lipogenesis and anabolic output control; and (4) cytoskeletal organization affecting cell migration and cancer progression. Variant effects are primarily through intronic variants that affect protein levels, particularly OSMR protein levels and blood cell measurements including hemoglobin concentration, red blood cell count, and reticulocyte fractions.

Evidence

- RICTOR component of mTORC2 transduces signals from growth factors to pathways involved in proliferation, cytoskeletal organization, lipogenesis and anabolic output
- RICTOR mTORC2 phosphorylates and activates AKT1, AKT2, AKT3, PKC family members, and SGK1
- RICTOR mTORC2 plays critical role in AKT1 activation by mediating phosphorylation of Ser-473, Ser-477, Thr-479, facilitating Thr-308 phosphorylation by PDPK1/PDK1

- RICTOR associated Reactome pathways include PI3K/AKT Signaling in Cancer, Constitutive Signaling by AKT1 E17K in Cancer
- RICTOR variants rs62359816, rs551454961, rs58330611 associated with OSMR protein levels with p-values 1e-120, 3e-77, 5e-88 respectively
- RICTOR variants rs769160064, rs200683443, rs1239259 associated with red blood cell count and hemoglobin concentration
- RICTOR variants rs13165709, rs12654031 associated with immature reticulocyte fraction measurements
- RICTOR predicted functions include regulation of TOR signaling, mRNA catabolic process, cellular response to oxidative stress
- RICTOR interacts with cancer-related proteins including MTOR, AKT1, BRCA1, MYC, EGFR, PRKCA through multiple experimental systems