PanelAppRex aggregates disease gene panels and facilitates sophisticated search

Dylan Lawless*[1]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.

March 27, 2025

Word count: 1342

Abstract

Motivation: Gene panel data provides critical insights into disease-gene correlations. However, aggregating and interrogating this diverse dataset can be challenging. PanelAppRex addresses this by first preparing a machine-readable aggregate and second by offering a sophisticated natural search interface that streamlines data exploration for both clinical and research applications.

Results: PanelAppRex aggregates gene panel data from source including CliVar, UniProt, and Genomics England (GE)'s PanelApp, including the approved panels used in the NHS National Genomic Test Directory and the 100,000 Genomes Project. It enables users to execute complex queries by gene names, phenotypes, disease groups and more, returning integrated datasets in multiple downloadable formats. Benchmarking demonstrates 93% - 100% accuracy, effectively simplifying variant discovery and interpretation to enhance workflow efficiency. The greatest benefit is the analysis ready format for bioinformatic integration.

Availability: `https://switzerlandomics.ch/services/panelAppRexAi/` (A standalone webpage will be substituted for publication version). The source code and data are accessible at `https://github.com/DylanLawless/PanelAppRex`. PanelAppRex is available under the MIT licence. The dataset is maintained for a minimum of two years following publication.

*Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch

Acronyms

RAG Retrieval-augmented generation 2

API Application Programming Interface 2

CSV comma-separated values 4

GE Genomics England 1

HTML HyperText Markup Language 3

PID Primary immunodeficiency 4

JACI Journal of Allergy and Clinical Immunology 4

PDF Portable Document Format 4

1 Introduction

Gene panels are pivotal for the diagnosis and interpretation of genetic disorders. Sources like GE's PanelApp and PanelApp Australia host comprehensive panels that support genomic testing (1). For instance, these are integral in the NHS and research projects such as the 100,000 Genomes Project(1). Despite its utility, manual panel selection and data aggregation remain labour intensive. PanelAppRex was developed to simplify this process by automating data retrieval from Application Programming Interface (API) sources and integrating the data into machine and user-friendly formats. We include the use of GE's PanelApp, ClinVar, and UniProt (1–3). The novel natural language-style search capability further streamlines the discovery of disease gene panels by allowing queries based on gene names, phenotypes, disease groups and additional key attributes which were supplemented with evidence-based Retrieval-augmented generation (RAG).

2 Materials and methods

2.1 Data

The PanelAppRex core model contained 58,592 entries consisting of 52 sets of annotations, including the gene name, disease-gene panel ID, diseases-related features, confidence measurements (4). Data from gnomAD v4 comprised 807,162 individuals, including 730,947 exomes and 76,215 genomes (5). This dataset provided 786,500,648 single nucleotide variants and 122,583,462 InDels, with variant type counts of 9,643,254 synonymous, 16,412,219 missense, 726,924 nonsense, 1,186,588 frameshift and 542,514 canonical splice site variants. ClinVar data were obtained from the variant summary dataset (this version: 16 March 2025) available from the NCBI FTP site, and included 6,845,091 entries, which were processed into 91,319 gene classification groups and a total of 38,983 gene classifications; for example, the gene A1BG contained four variants classified as likely benign and 102 total entries (2). For our analysis phase we also used dbNSFP which consisted of a number of annotations for 121,832,908 single nucleotide variants (6).

2.2 Implementation

PanelAppRex was implemented in R and integrates data from GE's PanelApp, ClinVar, and UniProt (1–3). It performed credentialed access to the API to retrieve all approved panels, merging them into two formats: a simplified version (Panel ID, Gene) and a complex version (including metadata such as confidence level, mode of inheritance, and disease information), and several metadata summary statistics. In addition, the tool incorporates a search module to execute complex user queries. The search functionality supports queries by gene names, phenotypes, disease names, disease groups, panel names, genomic locations and other identifiers. RAG was used to improve the natural queries in hidden states based on evidence about the disease and gene function by supplementing the data with additional sources including ClinVar, UniProt, etc.

2.3 Usage

Online, queries can be executed via the integrated search bar in our HyperText Markup Language (HTML) version, where a JavaScript function splits the query into individual terms and progressively filters rows - retaining only those that match all active terms while ignoring unmatched ones. This enables users to perform complex, partial matching queries (e.g. "paediatric RAG1 primary immunodeficiency skin disorder") to rapidly identify the panel most closely associated with their hypothesis on primary immunodeficiency and paediatric skin disorders.

Bioinformatically, users can import the provided, ready-for-use, datasets in TSV or Rds formats. Users are most likely to merge with their own omic data based on merging with gene/protein ID. The following code snippet, available in `minimal_example.R`, demonstrates how to load the data in R:

```
# TSV format
path_data <- "../data"
core_path <- paste0(path_data, "/PanelAppData_combined_core")
minimal_path <- paste0(path_data, "/PanelAppData_combined_minimal")

df_core <- read.table(
    file= paste0(core_path, ".tsv"),
    sep = "\t", header = TRUE)

df_minimal <- read.table(
    file= paste0(minimal_path, ".tsv"),
    sep = "\t", header = TRUE)

# Rds format
rds_path <- paste0(path_data, "/PanelAppData_combined_Rds")
df_core <- readRDS(file= rds_path)
```

3 Results

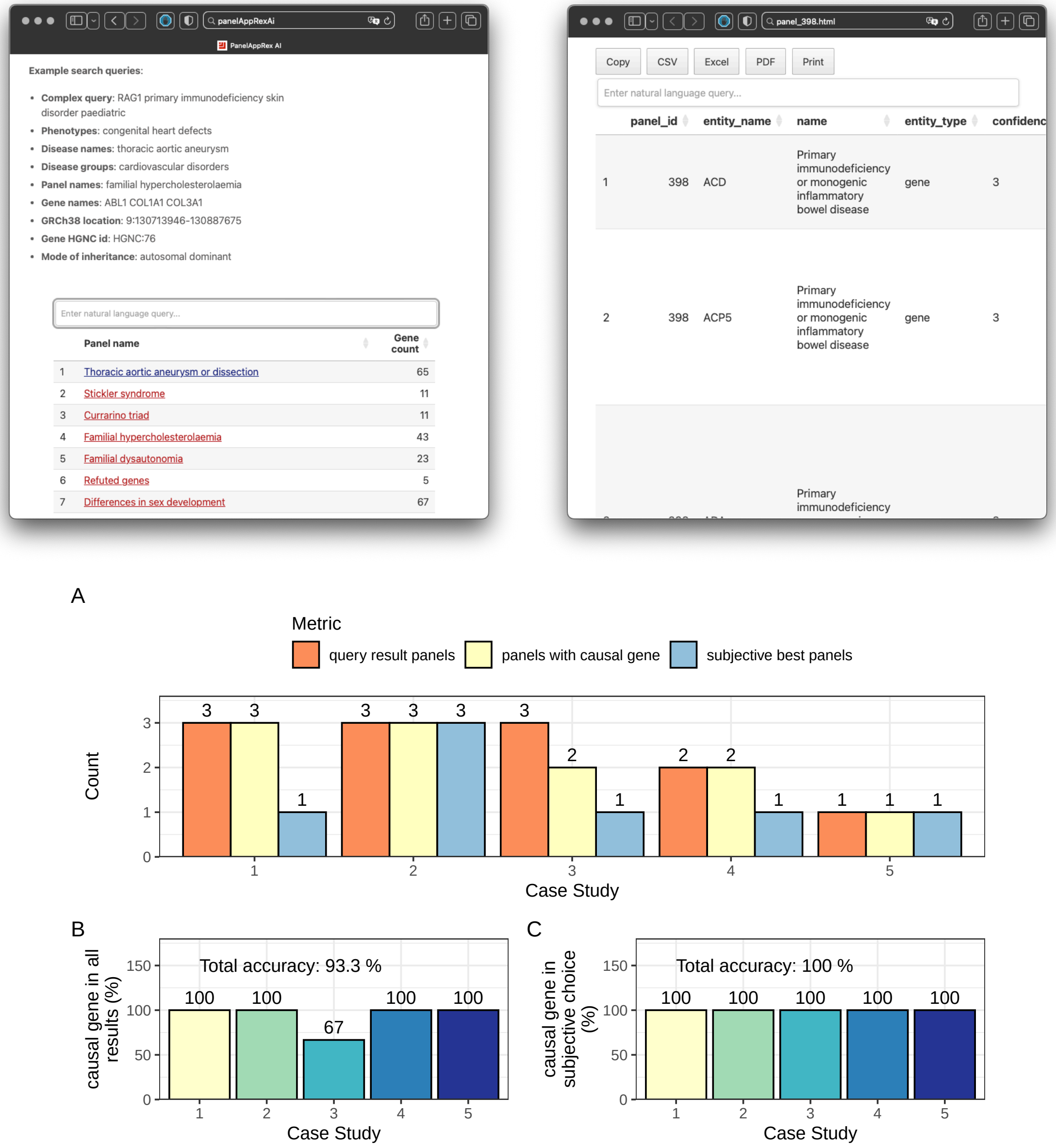
PanelAppRex successfully aggregates data, currently from 451 panels, and several genomics databases to offer a user-friendly search functionality (**Figure 1**). Users can retrieve results filtered by gene names, phenotypes, disease groups and other criteria. The system returns a table view with panel details and provides options for exporting results in comma-separated values (CSV), Excel, or Portable Document Format (PDF) formats. Bioinformatic uses may include virtual panels, for constructing prior odds, or for formal reporting on qualifying variants protocols.

3.1 Validation benchmarking

To mimic a clinician diagnosing a new disease such as Primary immunodeficiency (PID), we began by systematically selecting genetic diagnosis case studies from the current online catalogue from the Journal of Allergy and Clinical Immunology (JACI), using the first five results (7–11). We chose this source because our research centres on the genetic diagnosis of PID. The clinical background from these studies was used to construct keyword queries from patient features, simulating a naïve starting position for a clinician (full sources, queries, and results in **supplemental table 1**). We then tested whether our PanelAppRex tool could successfully retrieve panels that included the final causal gene reported in each case study.

The tool's query process retrieves gene panels using natural language-style input. For example, in the first case study on Hereditary Angioedema, the clinician reported suspecting that the condition was linked to "*SERPING1*, Factor XII, and edema" which we used as the query terms. Although the true causal gene, *F12*, was not mentioned, the inclusion of the related disease terms enabled the system to correctly retrieve the panel containing *F12* based on the hidden search knowledge.

In our evaluation, PanelAppRex returned panels in which the causal gene was present in 93% of all returned panels, with the user-selected best panel achieving a perfect accuracy of 100% (**Figure 1**). These metrics were derived by comparing the causal gene identified from the case study abstracts to the panels returned by our query, and by applying a subjective relevance measure to mimic intuitive panel selection - acknowledging that certain panels, such as the established PID gene panel, are inherently more reliable than broader, less specific panels. Overall, the results confirm that our approach accurately identifies the most relevant panels and effectively supports clinical decision-making in complex diagnostic scenarios.



Example search queries:

- **Complex query**: RAG1 primary immunodeficiency skin disorder paediatric
- **Phenotypes**: congenital heart defects
- **Disease names**: thoracic aortic aneurysm
- **Disease groups**: cardiovascular disorders
- **Panel names**: familial hypercholesterolaemia
- **Gene names**: ABL1 COL1A1 COL3A1
- **GRCh38 location**: 9:130713946-130887675
- **Gene HGNC id**: HGNC:76
- **Mode of inheritance**: autosomal dominant

Enter natural language query...

	Panel name	Gene count
1	Thoracic aortic aneurysm or dissection	65
2	Stickler syndrome	11
3	Currarino triad	11
4	Familial hypercholesterolaemia	43
5	Familial dysautonomia	23
6	Refuted genes	5
7	Differences in sex development	67

Enter natural language query...

	panel_id	entity_name	name	entity_type	confidenc
1	398	ACD	Primary immunodeficiency or monogenic inflammatory bowel disease	gene	3
2	398	ACP5	Primary immunodeficiency or monogenic inflammatory bowel disease	gene	3
			Primary immunodeficiency		

Figure 1: PanelAppRex interface displaying search results for a complex query. Top: screenshots showing the search interface, with the left panel displaying the full database before filtering and the right panel showing detailed results for a selected panel. Bottom: benchmark metrics are presented as follows: (A) for each of the 5 case studies, the total number of panels returned, the number of panels that included the causal gene, and the single best panel selected (always 1 by default); (B) the percentage of all returned panels that included the true causal gene; (C) the percentage of the single best panels that contained the true causal gene.

4 Summary

PanelAppRex offers a robust solution for aggregating and querying gene panel data. Its sophisticated search feature simplifies data exploration and enhances variant interpretation. Future work will focus on expanding the range of supported queries and integrating additional genomic data sources, further supporting the needs of clinicians and researchers.

Acknowledgements

We acknowledge GE for providing public access to the PanelApp data. The use of data from GE panelapp was licensed under the Apache License 2.0. The use of data from UniProt was licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). ClinVar asks its users who distribute or copy data to provide attribution to them as a data source in publications and websites (2).

Competing interest

We declare no competing interest.

References

[1] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple, Arianna Tucci, Helen Brittain, Anna De Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-019-0528-2. URL https://www.nature.com/articles/s41588-019-0528-2.

[2] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J Bradley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, January 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx1153. URL http://academic.oup.com/nar/article/46/D1/D1062/4641904.

[3] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Bowler-Barnett, and Others. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkae1010. URL https://academic.oup.com/nar/article/53/D1/D609/7902999.

[4] Dylan Lawless. PanelAppRex aggregates disease gene panels and facilitates sophisticated search. March 2025. doi: 10.1101/2025.03.20.25324319. URL http://medrxiv.org/lookup/doi/10.1101/2025.03.20.25324319.

[5] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[6] Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, 12(1):103, December 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00803-9. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9.

[7] Luisa Karla P. Arruda, Luana Delcaro, Priscila B. Botelho Palhas, Marina M. Dias, Valdair F. Muglia, Erick C. Castelli, Konrad Bork, and Adriana S. Moreno. Genetic Analysis As a Practical Tool to Diagnose Hereditary Angioedema with Normal C1 Inhibitor: A Case Report. *Journal of Allergy and Clinical Immunology*, 135(2):AB197, February 2015. ISSN 00916749. doi: 10.1016/j.jaci.2014.12.1578. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674914033594.

[8] Maeve A. McAleer, Elizabeth Pohler, Frances J.D. Smith, Neil J. Wilson, Christian Cole, Stuart MacGowan, Jennifer L. Koetsier, Lisa M. Godsel, Robert M. Harmon, Robert Gruber, Debra Crumrine, Peter M. Elias, Michael McDermott, Karina Butler, Annemarie Broderick, Ofer Sarig, Eli Sprecher, Kathleen J. Green, W.H. Irwin McLean, and Alan D. Irvine. Severe dermatitis, multiple allergies, and metabolic wasting syndrome caused by a novel mutation in the N-terminal plakin domain of desmoplakin. *Journal of Allergy and Clinical Immunology*, 136(5):1268–1276, November 2015. ISSN 00916749. doi: 10.1016/j.jaci.2015.05.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674915006533.

[9] Dorit Verhoeven, Dieneke Schonenberg-Meinema, Frédéric Ebstein, Jonas J. Papendorf, Paul A. Baars, Ester M.M. Van Leeuwen, Machiel H. Jansen, Arjan C. Lankester, Mirjam Van Der Burg, Sandrine Florquin, Saskia M. Maas, Silvana Van Koningsbruggen, Elke Krüger, J. Merlijn Van Den Berg, and Taco W. Kuijpers. Hematopoietic stem cell transplantation in a patient with

proteasome-associated autoinflammatory syndrome (PRAAS). *Journal of Allergy and Clinical Immunology*, 149(3):1120–1127.e8, March 2022. ISSN 00916749. doi: 10.1016/j.jaci.2021.07.039. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674921012446.

[10] Aude Magerus-Chatinet, Marie-Claude Stolzenberg, Nina Lanzarotti, Bénédicte Neven, Cécile Daussy, Capucine Picard, Nathalie Neveux, Mukesh Desai, Meghana Rao, Kanjaksha Ghosh, Manisha Madkaikar, Alain Fischer, and Frédéric Rieux-Laucat. Autoimmune lymphoproliferative syndrome caused by a homozygous null FAS ligand (FASLG) mutation. *Journal of Allergy and Clinical Immunology*, 131(2):486–490, February 2013. ISSN 00916749. doi: 10.1016/j.jaci.2012.06.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674912009645.

[11] Nigel Sharfe, Amit Nahum, Andrea Newell, Harjit Dadi, Bo Ngan, Sergio L. Pereira, Jo-Anne Herbrick, and Chaim M. Roifman. Fatal combined immunodeficiency associated with heterozygous mutation in STAT1. *Journal of Allergy and Clinical Immunology*, 133(3):807–817, March 2014. ISSN 00916749. doi: 10.1016/j.jaci.2013.09.032. URL https://linkinghub.elsevier.com/retrieve/pii/S0091674913014796.