RARe-SOURCE™: Integrated Bioinformatics Resource for Rare Diseases

Jason Cheung[1], Uma Mundunuri[2], PhD, Daniel Watson[2], Mohammad Alodadi[2], PhD, Erica Lyons[2], PhD, Anney Che[2], Gregory Tawa, PhD[1], London Toney[1], Cara Purdy[1], Richa Lomash[1], PhD, Stephanie Mounaud[1], Forbes Porter[3], MD, PhD, Sharie Haugabook[1], PhD, Elizabeth Ottinger[1], PhD.

[1]Therapeutic Development Branch, Division of Preclinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, 9800 Medical Center Drive, Rockville, MD, 20878
[2]Advanced Biomedical Computational Science, Frederick National Laboratory for Cancer Research, Frederick, MD, 21702
[3]Division of Translational Research, Eunice Kennedy Shriver National Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, 20892

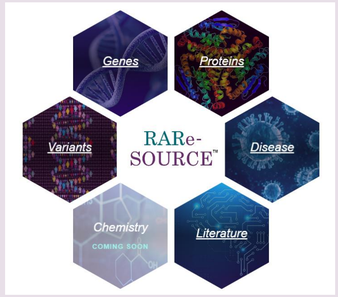
National Center for Advancing Translational Sciences

COLLABORATE. INNOVATE. ACCELERATE.

Introduction

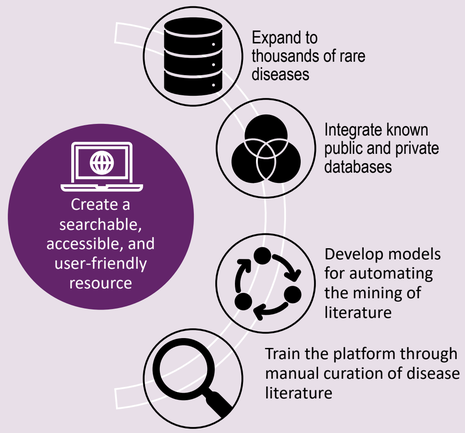
Rare disease data are fragmented and stored across different systems, platforms, and organizations. The ability to make connections between the diseases, their phenotypes, associated genes and related variants could unravel meaningful knowledge leading to the generation of novel hypotheses.

To harmonize and integrate these data sources, NCATS conceptualized RARe-SOURCE™, an integrated bioinformatics resource for rare diseases with the Advanced Biomedical Computational Science team at NCI-Frederick.



Scan Me!

The approach is to extract, annotate and integrate disease-associated data from reputable sources including the peer-reviewed scientific literature to help end-users make molecular associations and use available data to catalyze the discovery and development of treatments for rare diseases.

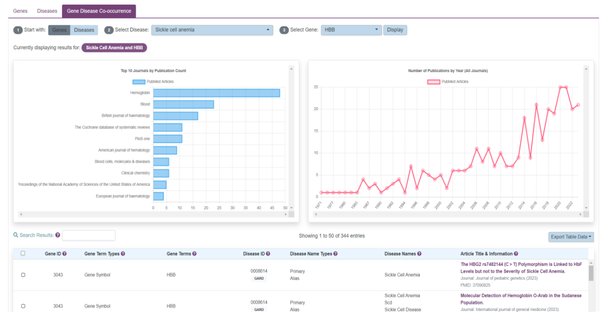


Rare Disease / Gene Information

RARe-SOURCE™ integrates multiple data sources to provide comprehensive information on rare diseases with known genetic associations. Disease ontologies were integrated from GARD[1] and Orphanet[2].

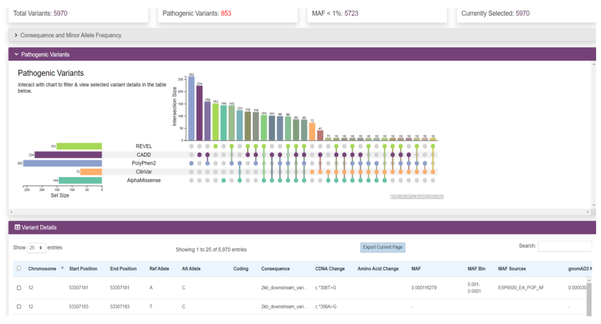
Literature AI

Generates results from mining rare disease and associated gene mentions in all MEDLINE™ by using transformer models. It has an added benefit of easing the search process for finding literature on diseases or genes of interest.



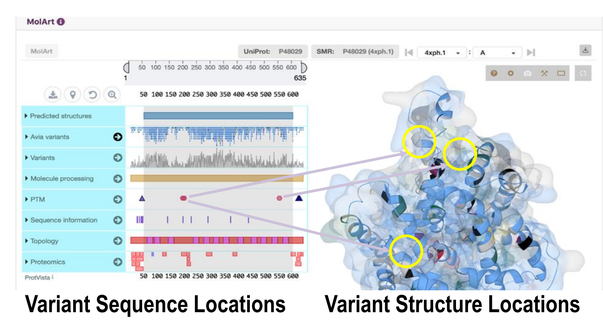
Gene Variant Annotation

Module includes millions of variants from public data sources such as gnomAD and ClinVar, annotated using OpenCRAVAT, for rare disease associated genes.



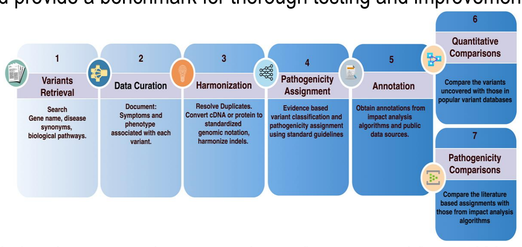
2D/3D Protein Structures

UniProt, MolArt, and Annotated Variants are integrated to illustrate the variant positions on the protein sequence and structure that can aid in functional impact analysis.



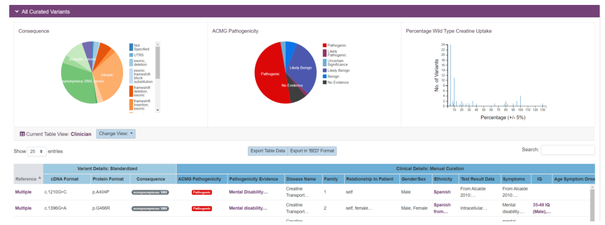
Manual Curated Variants

The manual curation analyses of literature related to rare diseases serve as a vital validation tool for our Literature AI language model and provide a benchmark for thorough testing and improvement.



This involves extensive comparison of current and future AI language models with manual curation and tools like PubMed and PubTator.

Manual curation of SLC6A8 for creatine transporter deficiency[3] provides a curated list of published variants.



The curated details are reported as a highly annotated dataset of variants with clinical context, functional details and interactive visualizations.

Future Directions

RARe-SOURCE™ is continuing to evolve by adding and integrating data and new features. The platform serves as a centralized information hub for rare diseases, to unlock novel insights into commonalities to identify new therapies to treat rare disorders.

Plans include integrating genotype-phenotype mapping, pathway associations, responses to chemical compounds, and disease correlations.

- Continue integrating public and private databases
- Utilize other artificial intelligence and machine learning tools to improve data collection
- Integrate phenotypic information into disease/gene literature mining
- Integrate epidemiological data
- Investigate chemical associations and commonalities in diseases/genes

Contact

https://raresource.nih.gov

rare-source@mail.nih.gov

Scan Me!

Notice: Feedback or reports of inaccuracies within the resource are welcomed to facilitate data accuracy and inclusivity.

References:

[1] Genetic and Rare Diseases Information Center (GARD). https://rarediseases.info.nih.gov/

[2] Orphanet: an online rare disease and orphan drug data base. https://www.orpha.net

[3] Lyons, E.L., Watson, D., Alodadi, M.S. et al. Rare disease variant curation from literature: assessing gaps with creatine transport deficiency in focus. BMC Genomics 24, 460 (2023). https://doi.org/10.1186/s12864-023-09561-5

Acknowledgements: The authors thank NCATS leadership for their support of this project.

Funding: This project is being supported by the Intramural Research Program of NCATS/NIH under Contract No. HHSN261201500003 through NCI. This project has been funded in part with Federal funds from the NCI/NIH/HHS under Contract No. 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of the trade names, commercial products, organizations imply endorsement by the U.S. Government.