RARe-SOURCE™: Integrated Bioinformatics Resource for Rare Diseases

Gregory J. Tawa[1], PhD, Uma Mudunuri[2], Daniel Watson[2], Mohammad Alodadi[2], PhD, Erica Lyons[2], PhD, Anney Che[2], PhD, Roy Satyaki[2], PhD, Richa Madan Lomash[1], PhD, London Toney[1], Cara Purdy[1], Stephanie Mounaud[1], Forbes Porter[3], MD, Sharie J. 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 Institute of Child Health and Human Development, NIH, Bethesda, MD, 20892.

National Center for Advancing Translational Sciences

COLLABORATE. INNOVATE. ACCELERATE.

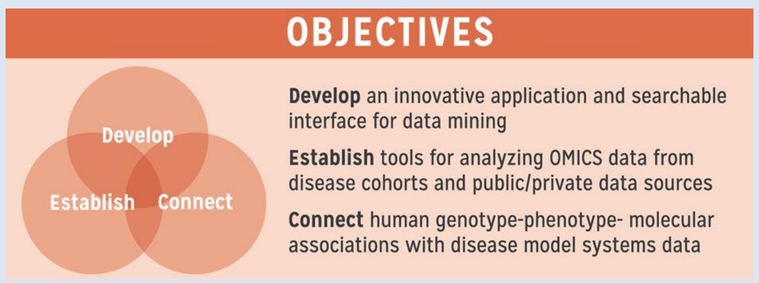
Introduction

Data is often stored in disparate systems, each with separate rules for access and use. There are also issues of integration and harmonization of data sets which are often based on different ontological schemes.

NCATS conceptualized Rare-SOURCE™ and then with Advanced Biomedical Computational Science at NCI-Frederick developed this Integrated Bioinformatic Resource for Rare Diseases.

The goals include development of query and analysis tools that span the combined data sets, allowing for connection of disparate concepts, and creation of new knowledge that will facilitate development of rare disease therapies.

To achieve these broad goals the objectives are to:

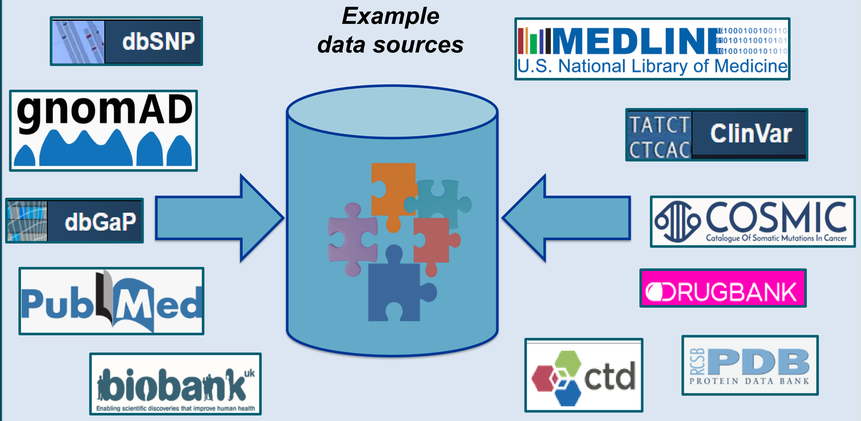


Challenges

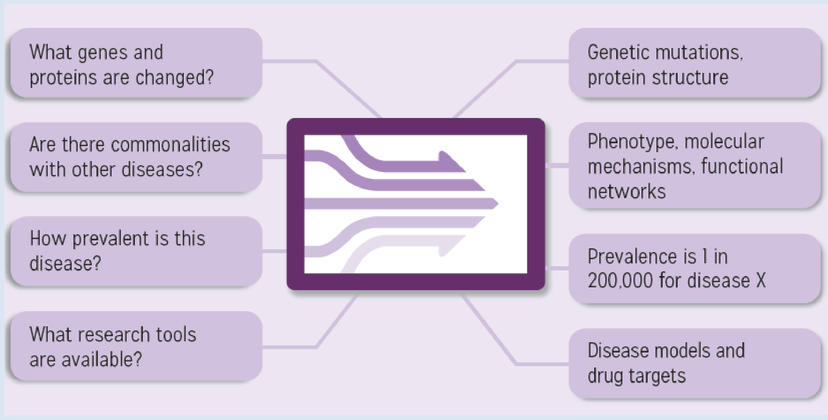
Data types span a wide range from molecular level details to genotypes, phenotypes, and disease consequences. Over 125 data sources were identified and are being integrated to create Rare-SOURCE™. This will make it possible for researchers to ask important questions about rare diseases.

Identify & Coordinate Disparate Data Sets

Integrate Data into Rare-SOURCE™

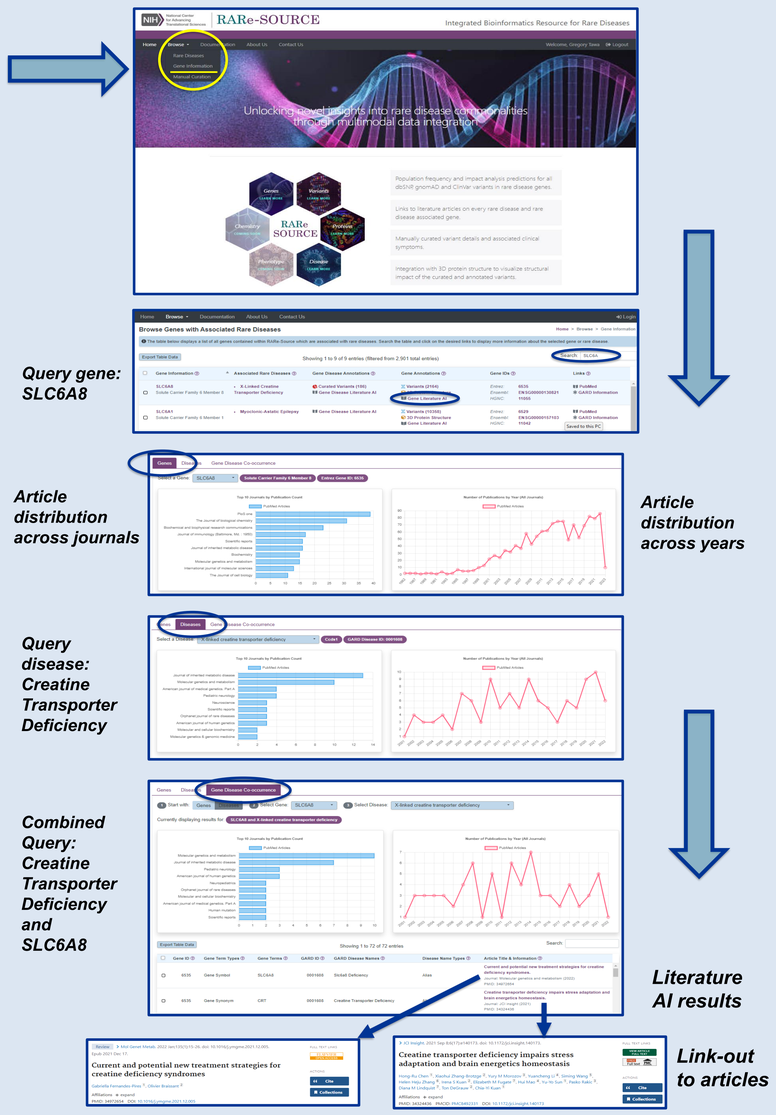


Data Platform



Literature Mining

The user interface allows users to begin an inquiry starting with a gene, a disease, or both. Artificial intelligence (AI) algorithms retrieve publications associated with query inputs. Not shown here, but deeper exploration can reveal molecular details of genetic variation and target protein structure.



Connect With Us

Website: https://raresource.nih.gov

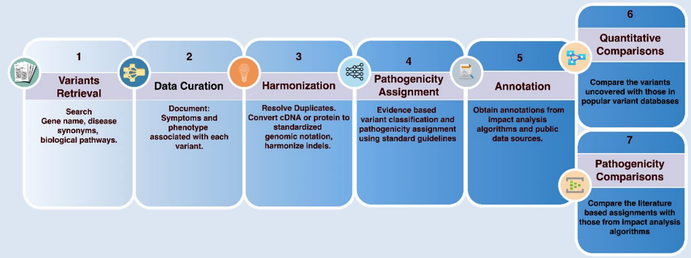
Email: rare-source@mail.nih.gov

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.

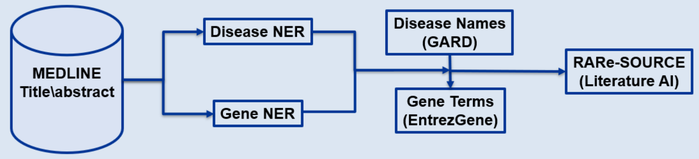
Variant Curation Pipeline

Manual curation analyses of the literature for pilot rare diseases serve as training sets for validation of natural language processing (NLP)-based AI models for gene-variants curation

Manual Curation Process



NLP/AI Algorithm Workflow



NLP-based AI will allow for scaling to analyze the literature for thousands of rare diseases

In Progress & Future Directions

ARCHITECTURE

- Manual curation
  - Data training sets
    - CTD
    - Farber Disease
- NLP for automated searches

RESEARCH

- Muscular Dystrophies
  - PhD Graduate Student (NCATS / NINDS)
- Research / Health Disparities
  - Highlighting rare diseases in systemically underrepresented populations

DATA

- Identify
- Outreach
  - Genetic / variant data
  - Phenotypic data
- Data Sharing Agreement
- Access/Connect/Share

MODULES

- Prevalence calculator
- Disease models
  - Pilot - Sickle Cell Pig NHS
- Disease Mechanisms/Biological Pathways/Drugs
- Equity component on preclinical research in rare diseases

Conclusions

Rare-SOURCE™ has been realized into a bioinformatics platform to help researches, clinicians, and rare disease patients/advocacy groups or other community stakeholders search for relevant information with a user-friendly interface for:

1) Linking to literature articles on rare diseases with genetic etiology and rare disease associated genes

2) Curation of gene/variant associations and predicted pathogenic significance

3) Integration with 3D protein structure to visualize the structural impact of the curated and annotated variants.

By connecting disparate data sources combined with AI literature mining, RARe-Source™ will be able to conduct systematic querying of rare disease data and accurately curate gene-disease variant associations. We anticipate that additional multimodal data integration will enable us to unlock novel insights into commonalities in rare diseases.