Retrieval, Reprocessing, Normalization and Sharing of Gene Expression and Airway Microbiome Data Sets to Facilitate AI/ML Analysis Studies of Bacterial Lung Infections

PI: Bruce A. Stanton, PhD

Presenter: Thomas H. Hampton, PhD

RESPIRE Team

Bruce Stanton

Matt Rogers

Kiyoshi Fukutani

Lily Taub

Sam Neff

RESPIRE: A community-based approach to big data

- I manage a bioinformatics core that specifically facilitates translational research in cystic fibrosis (CF) a genetic disease.

- Polymicrobial, antibiotic-resistant lung infections elicit immune responses that degrade lung function in CF.

- Being part of the CF community enables promote FAIR data principles and provide disease-focused analyses and software.

- RESPIRE enables CF researchers to quickly create data compendiums relevant to their research.

Challenges

- Researchers do not generally share count data

- Meta-data is highly variable in scope and format

- Designing and implementing an extensible system is much more difficult than a special purpose system

Progress

- https://respire.dartmouth.edu
- https://github.com/DartCF

	Gene Expression Module	Microbiome Module
Summary	Gene expression data related to lung diseases including asthma, pulmonary disease, cystic fibrosis, COPD, and idiopathic pulmonary fibrosis	Microbiome data related to lung diseases
Source of Metadata	NCBI (BioProject, GEO)	NCBI (BioProject)
Source of Data	refine.bio	SRA
Studies Identified	2,469	221
Studies Loaded	554	207
Samples Identified	72,813	50,021
Samples Loaded	24,142	40,619
Amount of Data	103 Gigabytes	.5 Gigabyte

Search RESPIRE Database

Select data module:

Gene Expression

Has data? ☑

Search terms

Enter one or more search terms

Select organism
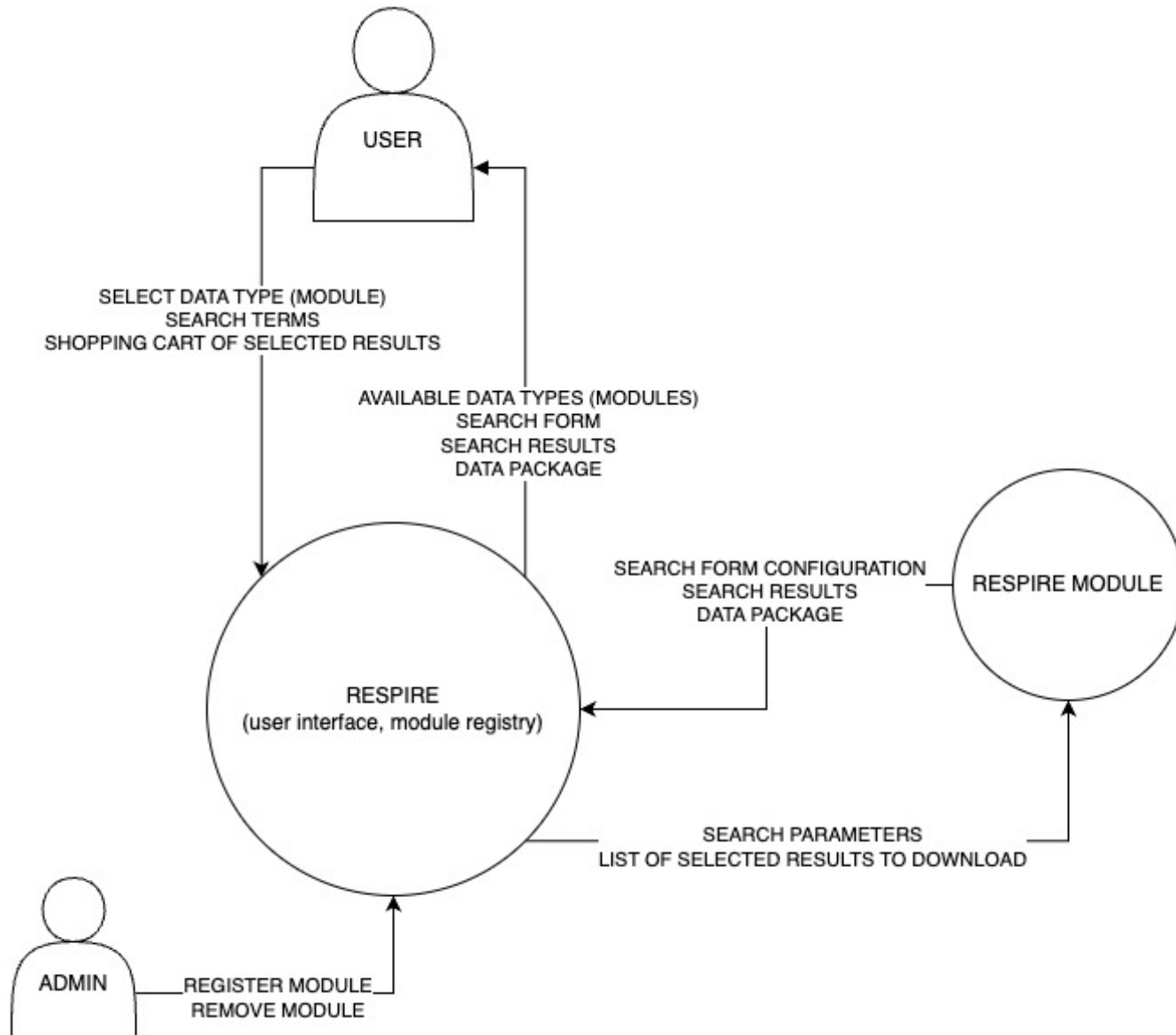
Pseudomonas aeruginosa

Select data profiling method

Expression profiling by array

Minimum Samples

50

Search

- Studies matching CF keywords are shown, further limited by key words
- Data sorted by number of samples
- Downloadable sets are self-consistent
- Single pipeline

- Extensible architecture
- Open source
- Documented
- Admin module

200

RESPIRE: Searchable, downloadable multi-omic data sets for lung disease research

T. Hampton[1], M. Rogers[2], S. Neff[3], B. Stanton[3]. [1]*Geisel School of Medicine, Dartmouth College, Hanover, NH;* [2]*Research Computing, Dartmouth College, Hanover, NH;* [3]*Department of Microbiology and Immunology, Geisel School of Medicine, Dartmouth College, Hanover, NH*

Background: The genomic revolution identified mutations in the CFTR gene as the underlying cause of CF, increasing understanding of disease etiology, identification of molecular targets, and ultimately highly effective modulator therapies that have vastly improved the lives of most people with this disease, but much work remains to be done. Modulators do not totally reverse the disease, and many people with CF (PwCF) have mutations that current modulators do not address. We believe that the next wave of innovation to address these issues will be driven in part by high-throughput transcriptomic, proteomic, metabolomic, and micro-biome data. Here we present theREpoSitory for Pulmonary expression data Reuse (RESPIRE), a platform designed to make publicly available multi-omic data easy for lung biologists to find, access, and reuse. Interoperable 'omics data will help drive basic science forward in CF and other lung diseases.

Dartmouth Scientific Partners
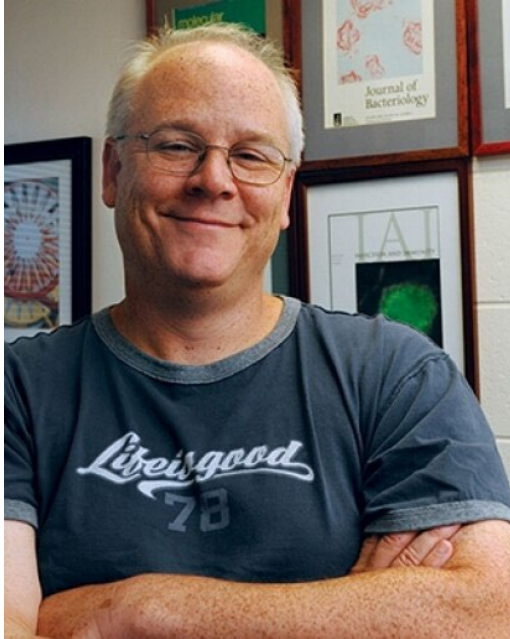

George O'Toole

Marta Rudzite

Bassam El Hafi

These collaborators use big public data sets to build model polymicrobial communities and study how less-familiar bacteria may drive patient outcomes in CF.

Ongoing Efforts

- Updating microbiome and transcriptomic sets

- Adding proteomic and metabolomic data

- Working with big pharma to roll out RESPIRE architecture in other disease indications

Acknowledgements

- Cystic Fibrosis Foundation
  - STANTO19G0
  - STANTO20PO
  - STANTO19R0
- NIH
  - P30-DK117469
  - R01HL151385
- Flatley Foundation