Using Data Systems to Increase Accessibility in Disease Specific Research

July 26, 2023

Megan Carnes, PhD
Alex Harding, MS
Ravi Mathur, PhD
Presenting Today

Megan Ulmer Carnes, PhD
Role: ME/CFS Network mPI
Technical Expertise: Genomics, Genetic Epidemiology, Microbiome, Bioinformatics

Alex Harding, MS
Role: map-systems Project Manager and Developer
Technical Expertise: Full-stack Web Applications, Data Visualization Dashboards, Cloud Infrastructure Management, Data Processing, and Software Architecture

Ravi Mathur, PhD
Role: map-systems Lead Bioinformaticist
Technical Expertise: Bioinformatics, Genetics, Metabolomics, Proteomics, Data Integration, Multi-Omics Analysis
Presentation Overview

- Introduction to ME/CFS: Megan Carnes
  - mapMECFS tool

- Open system framework: Alex Harding

- New portal development: Ravi Mathur
  - E.g., mapOA
ME/CFS – A complex, multi-factorial disease

**ME/CFS**
- Serious, long-term illness
- Affects many body systems
- Often limits people from doing their usual activities

**Symptoms**
- Severe fatigue
- Sleep problems
- Confined to bed
- Pain, dizziness, and difficulty with memory and cognition

**Post-Exertional Malaise (PEM)**
- Worsening of symptoms following mental or physical activity

**How many?**
- Institute of Medicine estimates 836,000 to 2.5 million Americans live with ME/CFS
- Most have not been diagnosed
One of the Data Management and Coordinating Center’s (DMCC) Goals: Build infrastructure to support secure sharing of data across a wide range of biological and clinical experiments.

ME/CFS-focused data repository

https://www.mapmecfs.org/
The Need for Data Integration

What’s needed in order to gain a better perspective

- Inventory of the range of data being collected
- An understanding of where other investigators are looking

Image from: http://melgen.org/multi-omics-approach/
<table>
<thead>
<tr>
<th>Researcher’s Challenge</th>
<th>The mapMECFS Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• I do not know where to go for ME/CFS-specific data.</td>
<td>• mapMECFS is a comprehensive repository of ME/CFS-related research data</td>
</tr>
<tr>
<td>• Sharing data is time consuming because my data files are complex and variable.</td>
<td>• Upload data using an easy step-by-step form with pre-populated fields</td>
</tr>
<tr>
<td></td>
<td>• Allows flexible file structures and data types</td>
</tr>
<tr>
<td>• I would like to compare my results other studies to draw biological conclusions.</td>
<td>• Contains custom search tools and curated ME/CFS literature to enable quick, cross-study comparisons</td>
</tr>
</tbody>
</table>
One Stop Shop for ME/CFS Research

Share New Data

- Create Dataset
- Data File
- Phenotype File
- Results File
- Supporting Files

* Required field

**Title:**
A descriptive title which should include the phenotype, data type, and data set. Examples: "ME/CFS case-control RNA expression study on Monocytic cells," "ME/CFS moderate vs. severe metabolomics study from the CFI cohort.

**URL:**
dataset/ cg_my-dataset

Description:

Search Existing Datasets

- Search by title, description, or molecules tagged in the data files or filter using the tags and metadata facets on the left.

- Example searches: TSPAN5, cg00000029, hsa-lent.7a, Fukuda, GSE128078

- 11 datasets found for "ATP"

- ME/CFS and QFS case-control RNA expression study GSE130353

- Raghavak et al. (2016) conducted a study on Chronic Fatigue Syndrome (CFS) and Q fever fatigue syndrome (QFS) where gene expression profiles were analyzed.

- Tags: case-control, GEO, QFS, RNA sequencing

Compare Study Results

Download Relevant Files

- Results File Explorer
- glucose

Example Searches: TSPAN5, cg00000029, hsa-lent.7a, IL-17

- D-glucose
  - Testing dataset for demo
  - Serum
  - Metabolomics
  - Wilcoxon Rank-Sum Test
  - 15.0 Control
  - 17.0 Patient

- D-glucose
  - Metabolic profiling of a ME/CFS syndrome discovery cohort
  - Serum
  - Metabolomics
  - Wilcoxon Rank-Sum Test
  - 15.0 Control
  - 17.0 Patient

- glucose
  - Prospective Biomarkers from Plasma
  - Plasma
  - Metabolomics
  - Wilcoxon Rank-Sum Test
  - 19.0 Control
  - 32.0 ME/CFS

- Bonferroni
  - 5.5105e-03
  - 1.0
Site Contents

61 PUBLIC DATASETS

>300 RESULT FILES

119 SITE USERS

https://www.mapmecfs.org/
mapMECFS Key Features: Search/Dataset

Search by title, description, or molecules tagged in the data files or filter using the tags and metadata facets on the left.

Example searches: TSPAN6, cg00000029, hsa-let-7a, Fukuda, GSE120078

7 datasets found for "IL6"

Order by: Relevance

ME/CFS case-control RNA expression study (GPL96) GSE14577
Gow et al. (2009) conducted a case-control study with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) on samples (8 ME/CFS and 7 healthy controls) on two...

Tags: GEO, has data file, has phenotype file

ME/CFS case-control RNA expression study (GPL97) GSE14577
Gow et al. (2009) conducted a case-control study with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) on all of 15 samples (8 ME/CFS and 7 healthy controls) on two...

Tags: GEO, has data file, has phenotype file

ME/CFS and IFS case-control RNA expression study on Monozygotic twins GSE1605
Byrne et al. (2009) conducted a study on myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) and idiopathic fatigue syndrome (IFS) using RNAseq.

Tags: GEO, has data file, has phenotype file

Files

Data File
- miRNA profiling data table
  - Type: TXT
  - Description: The normalized signal intensity of miRNA generated from the Ambion Bioarray V1.
  - Size: 111 KB

Phenotype File
- Phenotypes and covariates for Petty, et al.
  - Type: TXT
  - Description: Information on phenotypes and covariates for the Petty, et al. study.
  - Size: 21 KB

Supporting File
- GEO Public Data Use Agreement
  - Type: HTML
  - Description: Data use agreement for this public dataset which has been extracted from the...

Supporting File
- Data Generation Summary
  - Type: PDF
  - Description: The data generation summary for this study includes array information, study...
  - Size: 176 KB

Results File
- Example Result File
  - Type: TSV
  - Description: Calculated summary statistics from mapMECFS as an example result file.
  - Size: 32.5 KB

https://www.mapmecfs.org/
Results from experiment-specific analysis generally containing p-values or adjusted p-values

- Molecule column is searchable on mapMECFS
- Benefit is improved findability over a search in PubMed
- Custom tool (Result File Explorer) allows for comparisons across studies

Table 2 from manuscript

<table>
<thead>
<tr>
<th>Molecule</th>
<th>mean (CS)</th>
<th>sd (CS)</th>
<th>mean (CL)</th>
<th>sd (CL)</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL6</td>
<td>0.36</td>
<td>0.46</td>
<td>0.53</td>
<td>0.51</td>
<td>0.245</td>
</tr>
<tr>
<td>IL7</td>
<td>0.85</td>
<td>0.29</td>
<td>0.89</td>
<td>0.2</td>
<td>0.657</td>
</tr>
<tr>
<td>CXCL8</td>
<td>5.4</td>
<td>2.07</td>
<td>4.56</td>
<td>1.72</td>
<td>0.217</td>
</tr>
<tr>
<td>IL10</td>
<td>0.49</td>
<td>0.11</td>
<td>0.43</td>
<td>0.1</td>
<td>0.116</td>
</tr>
<tr>
<td>IL12p40</td>
<td>0.2</td>
<td>0.34</td>
<td>0.09</td>
<td>0.17</td>
<td>0.217</td>
</tr>
<tr>
<td>IL12p70</td>
<td>0.59</td>
<td>0.48</td>
<td>0.45</td>
<td>0.14</td>
<td>0.351</td>
</tr>
<tr>
<td>IL13</td>
<td>0.42</td>
<td>0.18</td>
<td>0.29</td>
<td>0.21</td>
<td>0.057</td>
</tr>
<tr>
<td>IL15</td>
<td>0.05</td>
<td>0.44</td>
<td>0.77</td>
<td>0.32</td>
<td>0.196</td>
</tr>
</tbody>
</table>

Results of t-tests of cytokine/chemokine levels comparing classical, short-duration ME/CFS cases to classical, long-duration ME/CFS cases (Hornig M, et al. Immune Transl Psychiatry. 2017)

https://www.mapmecfs.org/
mapMECFS Key Features: Easy Upload / Metadata

Create Dataset → Data File → Phenotype File → Results File → Supporting Files

Experimental Design
- Data Type: Methylation
- Assay: Illumina MethylationEPIC (850K) microarray
  Illumina Human Methylation 450k BeadChip
  Other
- Organism: (Dropdown)
- Measurement: (Dropdown)
  - B-values
  - Percent methylation (individual Cpg sites)
  - Percent methylation (global)
  - Raw intensity signals
  - Normalized intensity signals
  - Other
- Sample: (Dropdown)
- Phenotype: (Dropdown)

https://www.mapmecfs.org/
Custom Tools and Features to be Demoed

Molecule tagging
Uploaded data are tagged with known synonyms to improve searchability.

View Additional Search Terms

View Summary Statistics
Website Structure

- New users must be approved by NIH and agree to the DUA

- Uploaded data defaults to Private

- Public requests are reviewed for personally identifiable information before release
mapMECFS Demo
CKAN Framework

- **Comprehensive Knowledge Archive Network**
- Open-source data portal framework built in Python
  - Opinionated data storage in PostgreSQL
  - Robust CRUD (Create, Read, Update, Delete) structure with role-based user controls
- Data REST API
- Full-Stack application with HTML/Javascript frontend
- Extension architecture and community of extensions
- Authored by Open Knowledge Foundation, but recently moved to bilateral community stewardship
- Used for open data portals built by the US, Canadian, Australian governments to name a few
- Including data.gov!
CKAN allows users to upload any files or web URLs as Resources. Resources belong to Datasets, which are groups of Resources with shared metadata. Datasets belong to Organizations, which have members (users) with different levels of permissions.
Resources uploaded to CKAN can be processed for preview, including tables, charts, interactive visualization building tools, maps, and more.
CKAN provides a web-facing search interface.

These tools allow users to search **datasets** and **resources**, using a combination of text search, tags, and faceted search, similar to an Amazon shopping experience.
CKAN also provides an **HTTP REST API** for programmatic use of the data contained within, as well as for developers to interact with CKAN from advanced frontend tools for visualization and interfaces.

```python
ckan.logic.action.get.package_list(context, data_dict)
```

Return a list of the names of the site’s datasets (packages).

**Parameters:**
- `limit (int)` – if given, the list of datasets will be broken into pages of at most `limit` datasets per page and only one page will be returned at a time (optional)
- `offset (int)` – when `limit` is given, the offset to start returning packages from

**Return type:** list of strings
CKAN provides a robust extension API for software developers to add functionality to CKAN.

In the spirit of open-source software, many developers (including RTI) have chosen to make their work open-source and available.
RTI’s CKAN Extensions - Custom Tools

**User-Facing**

**SummaryStatistics**
- Generates descriptive statistics based on user supplied data
- Available on public GitHub

**SearchTerms**
- Data undergoes synonym matching to improve data findability
- Available on public GitHub

**ResultExplorer**
- Compiles results across datasets to compare and visualize data from multiple studies

**Backend**

**AdvancedAuth**
- Provides enhanced security features; protects data from unregistered users, initiates new user workflow, sharing public data request workflow
- Available on public GitHub

**AuditExplorer**
- Data access logs; enables quick response to data security and data quality issues

**QAChecker**
- Automated checking of datasets to easily identify datasets with processing errors
RTI’s New Data Integration Tool
MapMECFS AWS Infrastructure
Using AWS to replicate CKAN's docker-compose environment

Repositories on Github

ECR (container repository)

Local Machine

Deploy Process on Local Machine (automated):
- Clones repositories from Github
- Builds docker image
- Pushes image to ECR
- Triggers new tasks in ECS with ECR containers

EC2 Load Balancers
Staging Load Balancer
Production Load Balancer

Domain Names on Route53
staging.mapmeefs.org
mapmeefs.org

Port 443, Certificate in ACM
Port 5000
Port 5500

ECS Cluster
Staging
CKAN
redis
datapusher
Production
CKAN
redis

Tasks

Supervisor (Queue Workers)

EFS File Storage
RDS Aurora DB
Going from CKAN to mapMECFS

Human-Centered Design

Getting the Right Requirements

Small user group meetings

Getting the Requirements Right

Agile
With the extensions built and open-sourced and our infrastructure/automation abstracted to be more reusable, we can now build **new data portals** in CKAN, inheriting all the great work done by the mapMECFS team with minimal effort.
Welcome to mapOA, a data sharing portal created by the Integrative Omics Center for Opioid Addiction Research (IOMics-OA). The overarching goal of IOMics-OA is to accelerate the neurobiological understanding of opioid addiction (OA) and to identify biologically actionable drivers. The mapOA portal is designed to share opioid addiction-related data and results from IOMics-OA and others.

The site was created by RTI International with funding from NIH’s National Institute on Drug Abuse. Please visit the IOMics-OA website or email iomics-oa@rti.org for more information.
Bringing together multi-omic data from a variety of sources

With the variety of data, the map framework is being adapted to handle more data types (e.g., GWAS and integrated analyses), variety of analyses (e.g., meta-analysis and genomic SEM), and desired data security
Adapting map Framework

- Study design metadata is adapted to reflect Opioid Addiction, for example
  - Phenotype and Case definition
  - Comorbidities or OA characteristics
- Handling sample level and summary statistics data
- Analysis metadata is adapted to reflect new analyses including integrated analysis
  - Making relationships between datasets clear
- Data security is being adapted for eventual public consumption
- Development is coordinated with subject matter experts to design the desired system
FAIR Guiding Principles for scientific data management and stewardship

Findable  Accessible  Interoperable  Reusable


Want to learn more about mapMECFS?

Email the mapMECFS support team
mapMECFS@rti.org

Visit our website
https://www.mapmecfs.org/

delivering the promise of science for global good