mapMECFS: a portal to enhance data discovery across biological disciplines and collaborative sites

Ravi Mathur\*1, Megan Carnes\*1, Alexander Harding2, Amy Moore1, Ian Thomas2, Alex Giarrocco2, Michael Long2, Marcia Underwood2, Christopher Townsend2, Roman Ruiz-Esparza2, Quinn Barnette1, Linda Morris Brown1, Matthew Schu‡1

1Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA

2Center for Data Science, RTI International, Research Triangle Park, NC, USA

\*Co-first Authors

‡Corresponding Author – mschu@rti.org

**Additional File 1**

**Supplemental Methods**

mapMECFS Infrastructure

The comprehensive knowledge archive network1 framework includes a customizable containerization schema using Docker2 in conjunction with Amazon Web Services3 technologies to engineer a more performant, resilient, and affordable solution. Docker containers are minimal computing environments that can run on a personal computer, a dedicated server, or cloud computing. Docker containers, running computing environments, are created from Docker Images, container definitions, that allow the portal developers to define the computing environment, network connections, and source code needed for each component of the portal. These images can be tested locally and then pushed to the cloud computing instance and activated as containers running the most up-to-date version of the software in a stable and predictable fashion. mapMECFS’ database components are hosted using Amazon’s Relational Database Service Aurora4 database-as-a-service platform, which only incurs cost during times of usage and automatically scales to the size required by the contents of the database. User-uploaded files are stored using Amazon’s Elastic File System,5 a scalable and managed file system. By using these two services, we minimize cost of infrastructure and maintenance, as both services scale to usage for size and require no maintenance or management.

For the computation needed to run the application, mapMECFS utilizes Amazon’s Elastic Container Service (ECS)6 running on top of Elastic Compute Cloud (EC2).7 mapMECFS’ EC2 compute server functions as a provisioned virtual machine with adequate and scalable resources for handling loads from both web traffic and more resource-intensive asynchronous tasks such as *Calculated Summary Statistics* and *Synonym Tagging*. ECS allows developers to dynamically allocate resources throughout the containerized computation services, including the web server, a Redis cache for managing queues, and an instance of Apache Solr8 used for building search indexes. Additionally, mapMECFS leverages Amazon’s Cloudfront9 content delivery network and Route5310 domain name service to maintain availability across the world wide web.

mapMECFS Data Curation

 The Myalgic encephalomyelitis/chronic fatigue syndrome Research Network (MECFSnet)11 Data Management and Coordinating Center is continuously curating datasets from public repositories (e.g., Gene Expression Omnibus (GEO)12, MetaboLights13) and publicly available publications with open data reuse agreements. The datasets that have been curated as of March 2021 are shown in **Table S1**.

**Table S1:** *Public datasets curated into mapMECFS.*

|  |  |  |
| --- | --- | --- |
| **Publication** | **mapMECFS Data Type** | **Source\*** |
| Giloteaux et al.14, 15  | Cytokine Assay | MECFSnet Publication |
| Hornig et al.16 | Cytokine Assay | MECFSnet Publication |
| Hornig et al.17 | Cytokine Assay | MECFSnet Publication |
| Bouquet et al.18 | Gene Expression | GEO |
| Byrnes et al.19 | Gene Expression | GEO |
| Raijmakers et al.20 | Gene Expression | GEO |
| Gow et al.21 | Gene Expression | GEO |
| Armstrong et al.22 | Metabolomics | MetaboLights |
| Germain et al.23 | Metabolomics | MECFSnet Publication |
| Germain et al.24 | Metabolomics | MECFSnet Publication |
| Nagy-Szakal et al.25 | Metabolomics | MECFSnet Publication |
| Germain et al.26 | Metabolomics and Lipidomics | MECFSnet Publication |
| De Vega et al.27-29  | Methylation | GEO |
| Helliwell et al.30 | Methylation | Publication |
| Trivedi et al.31 | Methylation | GEO |
| Mandarano et al.32 | Microbiome | MECFSnet Publication |
| Nagy-Szakal et al.33 | Microbiome | MECFSnet Publication |
| Petty et al.34 | miRNA | GEO |
| Almenar-Perez et al.35 | miRNA | GEO |
| Billing-Ross et al.36 | mtDNA | MECFSnet Publication |
| Milivojevic et al.37 | Proteomics | MECFSnet Publication |

\*Curation is ongoing, we welcome ME/CFS research from all data sources.

Calculated Summary Statistics

For uploaded Data and Phenotype Files, mapMECFS generates a summary statistics file to characterize how the dataset measures compare between phenotype groups as annotated in the uploaded Phenotype file. A nonparametric Wilcoxon rank-sum test is calculated between groups within the study (e.g., between cases and controls). The summary statistics are automatically calculated for each molecule in the uploaded Data file. The summary statistics are processed asynchronously; therefore, they may not be immediately available after upload. The columns in the Summary Statistics file include (1) sample sizes in each group (labeled as “count”), (2) median value for each group, (3) standard deviation for each group, (4) Wilcoxon rank-sum test statistics (labeled as “Ranksum stat”), (4) Wilcoxon rank-sum p-value (labeled as “Ranksum p-value”), and (5) Wilcoxon rank-sum Bonferroni Corrected p-value (labeled as “Ranksum Bonf”).

**Figure S1:** *Schematic outlining the organizational structure of the mapMECFS portal, highlighting dataset privacy within Organizations.*



**Table S2:** *Synonym Tagging performed for each data type, which is driven by including a specific column name in the Data File (‘Required Data Column(s)’) and the process used databases to tag common alternate names for molecules. With the Synonym Tagging the search space for molecules expands beyond that provided in the Data File.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Type** | **Required Data Column(s)** | **Database used for Tagging** | **What is Searchable?** | **Example Searches** |
| Gene Expression | * Molecule
 | NCBI Gene (December 2018)38 | * Any entry from the Molecule data column
* Matching gene synonyms
 | * APOE
* ApoE4
 |
| Cytokine Assay | * Molecule
 | NCBI Gene (December 2018)38 | * Any entry from the Molecule data column
* Matching gene synonyms
 | * IL-17
* IL-17A
 |
| Metabolomics | * InChiKey39
* Molecule
* database\_identifier
 | N/A | * User input from any of the three required columns
 | * QTBSBXVTEAMEQO-UHFFFAOYSA-N
* Acetate
* CHEBI:15366
 |
| miRNA | * Molecule
 | miRBase (March 2019)40 | * Any entry from the Molecule data column
* Any miRNA related to the primary transcript.
* Any matching alias
 | * hsa-miR-198
* MIMAT0000228
 |
| Methylation | * Molecule
 | Illumina 450K (v.15017482\_v1-2) or Infinium MethylationEPIC (v-1-0-b4). | * Any entry from the Molecule data column
* Corresponding B37 coordinates (Chr:Pos)
 | * cg12045430
* 1:29407
 |

**References**

1. CKAN. <https://ckan.org/>.

2. Docker. <https://www.docker.com/>.

3. AWS. <https://aws.amazon.com/>.

4. Amazon Aurora. <https://aws.amazon.com/rds/aurora/>.

5. Amazon Elastic File System. <https://aws.amazon.com/efs/>.

6. Amazon Elastic Container Service. <https://aws.amazon.com/ecs/>.

7. Amazon EC2. <https://aws.amazon.com/ec2>.

8. Solr. <https://solr.apache.org/>.

9. Amazon CloudFront. <https://aws.amazon.com/cloudfront/>.

10. Amazon Route 53. <https://aws.amazon.com/route53/>.

11. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Network. <https://mecfs.rti.org/>.

12. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5.

13. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Research. 2019;48(D1):D440-D4.

14. Giloteaux L, Goodrich JK, Walters WA, Levine SM, Ley RE, Hanson MR. Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. Microbiome. 2016;4(1):30.

15. Giloteaux L, O’Neal A, Castro-Marrero J, Levine SM, Hanson MR. Cytokine profiling of extracellular vesicles isolated from plasma in myalgic encephalomyelitis/chronic fatigue syndrome: a pilot study. Journal of Translational Medicine. 2020;18(1):387.

16. Hornig M, Montoya JG, Klimas NG, Levine S, Felsenstein D, Bateman L, et al. Distinct plasma immune signatures in ME/CFS are present early in the course of illness. Science Advances. 2015;1.

17. Hornig M, Gottschalk CG, Eddy ML, Che X, Ukaigwe JE, Peterson DL, et al. Immune network analysis of cerebrospinal fluid in myalgic encephalomyelitis/chronic fatigue syndrome with atypical and classical presentations. Transl Psychiatry. 2017;7(4):e1080.

18. Bouquet J, Li T, Gardy JL, Kang X, Stevens S, Stevens J, et al. Whole blood human transcriptome and virome analysis of ME/CFS patients experiencing post-exertional malaise following cardiopulmonary exercise testing. PLoS One. 2019;14(3):e0212193.

19. Byrnes A, Jacks A, Dahlman-Wright K, Evengard B, Wright FA, Pedersen NL, et al. Gene expression in peripheral blood leukocytes in monozygotic twins discordant for chronic fatigue: no evidence of a biomarker. PLoS One. 2009;4(6):e5805.

20. Raijmakers RPH, Jansen AFM, Keijmel SP, Ter Horst R, Roerink ME, Novakovic B, et al. A possible role for mitochondrial-derived peptides humanin and MOTS-c in patients with Q fever fatigue syndrome and chronic fatigue syndrome. J Transl Med. 2019;17(1):157.

21. Gow JW, Hagan S, Herzyk P, Cannon C, Behan PO, Chaudhuri A. A gene signature for post-infectious chronic fatigue syndrome. BMC Med Genomics. 2009;2:38.

22. Armstrong CW, McGregor NR, Lewis DP, Butt HL, Gooley PR. Metabolic profiling reveals anomalous energy metabolism and oxidative stress pathways in chronic fatigue syndrome patients. Metabolomics. 2015;11(6):1626-39.

23. Germain A, Ruppert D, Levine SM, Hanson MR. Prospective Biomarkers from Plasma Metabolomics of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Implicate Redox Imbalance in Disease Symptomatology. Metabolites. 2018;8(4).

24. Germain A, Ruppert D, Levine SM, Hanson MR. Metabolic profiling of a myalgic encephalomyelitis/chronic fatigue syndrome discovery cohort reveals disturbances in fatty acid and lipid metabolism. Mol Biosyst. 2017;13(2):371-9.

25. Nagy-Szakal D, Barupal DK, Lee B, Che X, Williams BL, Kahn EJR, et al. Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. Sci Rep. 2018;8(1):10056.

26. Germain A, Barupal DK, Levine SM, Hanson MR. Comprehensive Circulatory Metabolomics in ME/CFS Reveals Disrupted Metabolism of Acyl Lipids and Steroids. Metabolites. 2020;10(1).

27. de Vega W, Erdman L, Vernon SD, Goldenberg A, McGowan PO. . Integration of DNA methylation & health scores identifies subtypes in myalgic encephalomyelitis/chronic fatigue syndrome. Epigenomics. 2018;10(5):539-57.

28. de Vega WC, Herrera S, Vernon SD, McGowan PO. Epigenetic modifications and glucocorticoid sensitivity in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS). BMC Med Genomics. 2017;10(1):11.

29. de Vega WC, Vernon SD, McGowan PO. DNA methylation modifications associated with chronic fatigue syndrome. PLoS One. 2014;9(8):e104757.

30. Helliwell AM, Sweetman EC, Stockwell PA, Edgar CD, Chatterjee A, Tate WP. Changes in DNA methylation profiles of myalgic encephalomyelitis/chronic fatigue syndrome patients reflect systemic dysfunctions. Clinical Epigenetics. 2020;12(1):167.

31. Trivedi MS, Oltra E, Sarria L, Rose N, Beljanski V, Fletcher MA, et al. Identification of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome-associated DNA methylation patterns. PLoS One. 2018;13(7):e0201066.

32. Mandarano AH, Giloteaux L, Keller BA, Levine SM, Hanson MR. Eukaryotes in the gut microbiota in myalgic encephalomyelitis/chronic fatigue syndrome. PeerJ. 2018;6:e4282.

33. Nagy-Szakal D, Williams BL, Mishra N, Che X, Lee B, Bateman L, et al. Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. Microbiome. 2017;5(1):44.

34. Petty RD, McCarthy NE, Le Dieu R, Kerr JR. MicroRNAs hsa-miR-99b, hsa-miR-330, hsa-miR-126 and hsa-miR-30c: Potential Diagnostic Biomarkers in Natural Killer (NK) Cells of Patients with Chronic Fatigue Syndrome (CFS)/ Myalgic Encephalomyelitis (ME). PLoS One. 2016;11(3):e0150904.

35. Almenar-Perez E, Sarria L, Nathanson L, Oltra E. Assessing diagnostic value of microRNAs from peripheral blood mononuclear cells and extracellular vesicles in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Sci Rep. 2020;10(1):2064.

36. Billing-Ross P, Germain A, Ye K, Keinan A, Gu Z, Hanson MR. Mitochondrial DNA variants correlate with symptoms in myalgic encephalomyelitis/chronic fatigue syndrome. J Transl Med. 2016;14:19.

37. Milivojevic M, Che X, Bateman L, Cheng A, Garcia BA, Hornig M, et al. Plasma proteomic profiling suggests an association between antigen driven clonal B cell expansion and ME/CFS. PLoS One. 2020;15(7):e0236148.

38. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005;33(Database issue):D54-8.

39. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform. 2015;7:23.

40. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155-D62.