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

Ravi Mathur (rmathur@rti.org)
RTI International https://orcid.org/0000-0001-6785-0032

Megan U. Carnes
RTI International

Alexander Harding
RTI International

Amy Moore
RTI International

Ian Thomas
RTI International

Alex Giarrocco
RTI International

Michael Long
RTI International

Marcia Underwood
RTI International

Christopher Townsend
RTI International

Roman Ruiz-Esparza
RTI International

Quinn Barnette
RTI International

Linda Morris Brown
RTI International

Matthew Schu
RTI International

Research Article

Keywords: Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), Data Sharing Portal, Comprehensive Knowledge Archive Network (CKAN), Data Integration, Multi-Omics

Posted Date: May 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-522847/v1

Abstract

Background

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a debilitating disease which involves multiple body systems (e.g., immune, nervous, digestive, circulatory) and research domains (e.g., immunology, metabolomics, the gut microbiome, genomics, neurology). Despite several decades of research, there are no established ME/CFS biomarkers available to diagnose and treat ME/CFS. Sharing data and integrating findings across these domains is essential to advance understanding of this complex disease by revealing diagnostic biomarkers and facilitating discovery of novel effective therapies.

Methods

The National Institutes of Health funded the development of a data sharing portal to support collaborative efforts among an initial group of three funded research centers. This was subsequently expanded to include the global ME/CFS research community. Using the open-source comprehensive knowledge archive network (CKAN) framework as the base, the ME/CFS Data Management and Coordinating Center developed targeted metadata collection, smart search capabilities, and domain-agnostic data integration to support data findability and reusability while reducing the barriers to sustainable data sharing.

Results

We designed the mapMECFS data portal to facilitate data sharing and integration by allowing ME/CFS researchers to browse, share, compare, and download molecular datasets from within one data repository. At the time of publication, mapMECFS contains data curated from public data repositories, peer-reviewed publications, and current ME/CFS network researchers.

Conclusions

mapMECFS is a disease-specific data portal to improve data sharing and collaboration among ME/CFS researchers around the world. mapMECFS is accessible to the broader research community with registration. Further development is ongoing to include novel systems biology and data integration methods.

Background

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a complex, debilitating disease[1] estimated to affect as many as 2.5 million Americans.[2] Affected individuals often have incapacitating fatigue, nonrefreshing sleep or other sleep difficulties, and cognitive impairment that may leave them unable to leave the house or bed. The disease is characterized by the worsening of symptoms following even minor physical or mental exertion, known as post-exertional malaise.[3] Although the underlying disease etiology remains unknown,[1, 4] there is evidence that multiple body systems are involved. When comparing ME/CFS cases to controls, researchers have observed differences in the immune system,[5, 6] blood metabolites,[7–12] the gut microbiome,[13–15] and mitochondrial DNA genetic variants.[16] Integrating findings across these domains promises to reveal a more complete picture of the disease, thereby detecting diagnostic biomarkers and facilitating discovery of novel effective therapies.

To support data sharing, the ME/CFS Research Network[17] comprising three Collaborative Research Centers and a Data Management and Coordinating Center (DMCC) was funded in 2017 by multiple National Institutes of Health (NIH) Institutes, Offices, and Centers, including the National Institute of Neurological Disorders and Stroke and the National Institute of Allergy and Infectious Diseases to encourage collaborative research to lead to better diagnosis and treatment for ME/CFS. One of the goals of the DMCC is to help ME/CFS researchers discover new disease insights by promoting data sharing. To this end, we developed the mapMECFS[18] data portal, which is built on a flexible database structure and optimized to handle multiple data types (e.g., gene expression, methylation, metabolomics, cytokine measures, proteomics, microbiome, survey/questionnaire). This portal contains specialized features to support cross-disciplinary and cross-study ME/CFS research generated from multiple research domains including cascading forms to collect key study metadata, intuitive dataset filtering, and smart search capabilities with synonym tagging with embedded mapping of synonymous feature terminologies.

Methods

Website framework

mapMECFS is an intuitive data portal built on the comprehensive knowledge archive network (CKAN) framework, an open-source tool designed to support data storage and sharing.[19] The portal includes an ecosystem of custom plugins hosted on a novel infrastructure built on Amazon Web Services technologies.[20]

The DMCC has created custom plugins for mapMECFS, including expansion of CKAN's default metadata structure for datasets and files, search formatting, mapMECFS-specific features and tools, and authentication changes to facilitate website access and data privacy (see **Additional File 1: Supplementary Methods**). The DMCC has made significant contributions back to the open-source CKAN community[21] including fixes for outstanding issues, enhancements to infrastructure supporting extension development, security, authorization, and reusability of the infrastructure.

mapMECFS data contents and processing

mapMECFS was designed to store de-identified demographic, survey, and health data coupled with molecular data (e.g., transcriptomics, metabolomics, methylation). The DMCC has curated ME/CFS data with open-access publications[7–16]; gene expression, methylation, and micro-RNA (miRNA) datasets from the Gene Expression Omnibus[22]; and metabolomic data from MetaboLights.[23] Active curation by the DMCC is ongoing (see **Additional File 1: Supplemental Methods** and **Table S1**).

Metadata, such as dataset title, description, tags, cohort selection, and case definition, are captured during the upload process for all datasets. The dynamic upload process prompts users for other key metadata based on the data type. Thus, metadata prompted for Gene Expression datasets will be different than metadata prompts for DNA Methylation experiments. Metadata collection is performed using prepopulated, easy-to-use drop-down menus to describe the experimental assay and data measurement unit with supplemental open-ended text boxes. Uploaded data are categorized by tags, which are suggested by natural language processing in real time based on a description provided by the user. Users can select the most appropriate suggested tags or supply their own as free text. These tags are used to filter and sort datasets to enable easy findability.

Once uploaded, data are processed with the custom features, *Synonym Tagging* and *Calculated Summary Statistics*. *Synonym Tagging* is an automated, backend feature of mapMECFS that labels molecules with known synonyms to enhance the searchability of analytes on the portal. By tagging both the indicated annotation and all recognized synonyms, mapMECFS extends the search space for each entered query. Synonyms are assigned based on well-established databases: National Center for Biotechnology Information's Entrez[24] for gene expression data; InChIKey,[25] ChEBI[26] ID, and HMDB[27] ID for metabolomic data; miRbase[28] for miRNA data; and Illumina manifest files for methylation Illumina array data. For transparency, the results of the *Synonym Tagging* are available on the dataset page (**Additional File 3: Figure S3**). If the uploaded dataset contains both a data file with sample-level molecule values and a phenotype file with participant-level variables (e.g., case-controls status), the Calculated *Summary Statistics* tool will calculate and display a set of summary statistics. These two-group comparison tests are performed using the nonparametric Wilcoxon Rank Sum test, which is calculated independently for each molecule compared between each phenotypic group (e.g., case vs. control, multiple subtypes). Further details are described in **Additional File 1: Supplemental Methods**.

mapMECFS access and site organization

mapMECFS is accessible to the broader research community at https://www.mapmecfs.org.[18] However, to obtain full access to the mapMECFS portal, new users must register for an account, provide a brief description of how they will use the data on the portal, and agree to the data use agreement. Account registration must be approved by the NIH data access committee. Upon approval, a user account will be created, granting the user permission to explore and upload data.

mapMECFS site users are grouped together within an "Organization," which designates the institute, research center, or individual research lab that users and the datasets they upload are associated with (**Additional File 1: Figure S1**). A user must be part of an Organization to be allowed to upload data.

Independent or unaffiliated users are assigned to an Organization where they are the only member. Datasets are designated as either public (data are available to all mapMECFS site users) or private (data are only available to users within the organization). Each user account has a defined role within an organization. Organization members can view private datasets within the organization on a read-only basis. Organization editors have all the abilities of members in addition to the ability to create and edit datasets they created. Editors cannot edit datasets by other editors in their Organization. Organization administrators have all the abilities of an editor in addition to the ability to edit any dataset within the Organization, change dataset accessibility (public vs. private), and change users' roles within their Organization. Requesting a dataset be set to "Public" initiates a review by the mapMECFS support teams to confirm that the data being made public adhere to our policies for de-identifying subjects and protecting personally identifiable information. Once the review is complete, DMCC administrators will push the dataset to a Public setting.

Results

We created mapMECFS to facilitate sharing of ME/CFS data among the broader research community. To expedite that process, we populated the portal with current publications and publicly available data. An overview of the portal is shown in **Figure 1**. Registered site users can upload their own de-identified primary research data and research results and share with other approved site users by creating a dataset and uploading associated files.

Figure 1: *mapMECFS website overview.* *Data flow chart for mapMECFS. The user uploads files (data, phenotype, results and supporting files) to a Dataset along with metadata. User uploaded files All datafiles are access protected to public or private, depending on the user's preference. mapMECFS processes the data to generate summary statistics and conduct synonym tagging based on the data type. The user can search available data along with viewing, filtering, or downloading files.*

The definition of a dataset is flexible; it can contain one or many relevant file types (data file, phenotype file, results file, supporting files), as described in **Table 1**. To take advantage of the *Synonym Tagging* and *Calculated Summary Statistic* features (described in the **Methods** section), the data, phenotype, and results files must follow the specified file format requirements described in **Additional File 2: Figure S2**. A dataset can contain an unlimited number of results and supporting files, thus enabling the sharing of standard operating procedures, external links (e.g., publication, sequencing read archive, data availability), and results from two or more analyses. The data upload process is flexible, allowing for multiple file formats. We encourage researchers to share datasets with other mapMECFS site users by making their datasets Public (i.e., viewable to all approved mapMECFS users). The process of

approving public data requests will be conducted in accordance with the mapMECFS policies (see website and **Methods** section).

Table 1. *File types uploaded to mapMECFS, including data file (e.g., processed data), phenotype file (e.g., clinical data), results file (e.g., summary statistics), and support file (e.g., link to publication).*

File Type	*Description*	*Examples*
Data File	Processed data containing sample-level values as columns and molecules as rows. The header for each column should match the participant ID in the phenotype file. Only one data file can be uploaded per dataset.	• Gene expression counts • Methylation signal intensities • Metabolomics mass spectrometry peak heights
Phenotype File	Subject-level clinical values with participant ID matching that in the data file. Only one phenotype file can be uploaded per dataset.	• Case-control status • Age • Sex • Relevant covariates
Results File	Summary statistics and other analysis output generated by the user, with statistics reported for each molecule.	• Wilcoxon-rank sum summary statistics with p-value and adjusted p-value included as columns
Supporting File	Additional documentation of experimental procedures or supporting material. Supporting documentation is recommended to provide users with a better understanding of the experiment generating the dataset.	• Standard operating procedures describing the dataset generation in more detail • Hyperlinks to publications using the data included in the dataset • Data dictionary

We optimized the mapMECFS search functionality to maximize data discovery. The search feature recognizes user-specified terms describing multiple aspects of a dataset, including a dataset name, descriptions, key metadata, data file contents, and data synonyms generated from the *Synonym Tagging* tool (**Methods** and **Additional File 1: Table S2**). For example, if a researcher is interested in the cytokine interleukin 17, searches of “IL-17” or “IL-17A” will

return all datasets containing the desired molecule, effectively standardizing terminology from different research domains and ontologies.

mapMECFS contains a customized data search tool, the *Results File Explorer*, to facilitate cross-dataset searches. This tool is designed to allow users to quickly evaluate the reproducibility of a given result across multiple studies by comparing results to other analyses or subset analyses. **Additional File 4: Figure S4** shows an example search for the metabolite 4-hydroxyglutamate. This amino acid, part of the glutamate metabolism pathway, has substantial implications in brain function[7] and has been shown to be elevated in ME/CFS patients.[7] The results shown in **Additional File 4: Figure S4** contain three separate tables, “Data Files and Calculated Summary Statistics,” “Results Files,” and “Other Datasets.” “Data Files and Calculated Summary Statistics” contains search results only from the mapMECFS *Calculated Summary Statistics*. The “Results Files” table contains search results only from the user-uploaded results files. The “Other Datasets” table contains search results from other elements of the dataset, including the title, description, or metadata. The tables contain key metadata so researchers can view analytic results while comparing analysis endpoints and applied methods. With the *Results File Explorer* tool and the mapMECFS search functionality users can identify datasets to validate their findings and identify datasets to integrate with data they have collected. This process enhances the collaborative nature of ME/CFS research.

Discussion

To promote international sharing of ME/CFS data across researchers and repositories, we created mapMECFS, a ME/CFS domain-specific data repository that conforms to NIH’s Findable, Accessible, Interoperable, and Reusable (FAIR)[29] Guiding Principles and its strategic plan for data science.[30] mapMECFS achieves this by (1) being findable with persistent identifiers for each table within a dataset and the dataset itself and providing rich metadata, (2) making all data and metadata accessible via both an intuitive user interface and application programming interface, (3) providing data and metadata that are interoperable with vocabulary and language which are common in the field, and (4) using clear reuse license and data provenance. The mapMECFS portal is developed with a scalable infrastructure that delivers scientific impact to the research community, supports good data management practices, and actively engages the user community.

mapMECFS use cases may include (1) identifying new collaborators by identifying researchers working on a specific data type or by browsing descriptions of individual researchers’ interests, (2) validating a research finding by searching for a molecule of interest (in the Dataset explorer or *Results File Explorer*) to discover independent datasets containing that molecule, (3) meeting a funder or journal’s requirements of

data sharing by uploading new experimental data or summary statistics, and (4) increasing statistical power by identifying an appropriate dataset for meta-analysis or data integration using the captured metadata, dataset description, and other support documentation. The DMCC plans to develop improved methods of data integration by standardizing sample identifiers and clinical data capture to support systems biology and data integration approaches for ME/CFS research. Investigating multi-omics may help with identification for disease subtyping, diagnosis, and predictive outcome.[31]

Conclusions

mapMECFS18 is available to the broader research community with registration as described above. To facilitate sharing of MECFS data, we encourage new users to consider mapMECFS for their research needs. This unique, domain-specific data repository was designed considering NIH's FAIR Guiding Principles and strategic plan for data science. We encourage mapMECFS users to provide feedback and to request new features. mapMECFS is continuously expanding the types of data included on the site and welcomes feedback on the types of data researchers would like to share.

Abbreviations

ChEBI
Chemical Entities of Biological Interest
CKAN
comprehensive knowledge archive network
DMCC
Data Management and Coordinating Center
FAIR Principles
Findable, Accessible, Interoperable, Reproducible Principles
HMDB
Human Metabolomics Database
InChIKey
International Chemical Identifier Key
ME/CFS
Myalgic encephalomyelitis/chronic fatigue syndrome
miRNA
micro-RNA
NIH
National Institutes of Health

Declarations

Ethics approval and consent to participate: N/A

Consent for publication: N/A

Availability of data and materials: The mapMECFS data portal is open to the research community at https://www.mapmecfs.org/

Competing interests: None to declare

Funding: Research reported in this publication was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number U24NS105535. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' contributions: RM and MC led the writing of this work. MC and AH are the Bioinformatics and development leads for mapMECFS. RM, MC, AH, AM, IT, AG, ML, MU, CT, RRE, and QB are technical staff working on the development of mapMECFS. LMB and MS are the co-PIs on the ME/CFS DMCC and advised the development of mapMECFS. All authors edited and reviewed the manuscript.

Acknowledgements:

The authors would like to thank all members of the ME/CFS network including:

Andrew Breeden[1,2], Joseph J. Breen[3], JoNita Cox[4], Maureen Hanson[5], Keith LeGrow[4], Ian Lipkin[6], Taya McMillan[7], Rebecca B. McNeil[4], Alain Moreau[8], Callie Riggs[4], Derya Unutmaz[9], Lawrence Whitley[4], Vicky Whittemore[1], Laura Elizabeth Wiener[4]

[1]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA.

[2]Currently at the National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA.

[3]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA.

[4]Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA.

[5]Center Director, Cornell NIH Collaborative ME/CFS Center, Cornell University, Ithaca, NY, USA.

[6]Center Director, Columbia Center for Solutions for ME/CFS, Columbia University, New York, NY, USA.

[7]Center for Communication Science, RTI International, Research Triangle Park, NC, USA.

[8]Center Director, Interdisciplinary Canadian Collaborative Myalgic Encephalomyelitis (ICanCME), ICanCME Research Network, Montréal, Quebec, Canada.

[9]Center Director, Jackson laboratory ME/CFS Collaborative Research Center, Jackson Laboratory for Genomic Medicine, Farmington, CT, USA.

References

1. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. https://www.cdc.gov/me-cfs/index.html.
2. Medicine Io. Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness. Washington, DC: The National Academies Press; 2015. 304 p.
3. Carruthers BM, Jain AK, De Meirleir KL, Peterson DL, Klimas NG, Lerner AM, et al. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Journal of Chronic Fatigue Syndrome. 2003;11(1):7–115.
4. Cortes Rivera M, Mastronardi C, Silva-Aldana CT, Arcos-Burgos M, Lidbury BA. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Comprehensive Review. Diagnostics (Basel). 2019;9(3).
5. 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.
6. 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.
7. 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).
8. 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.
9. 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).
10. Mandarano AH, Maya J, Giloteaux L, Peterson DL, Maynard M, Gottschalk CG, et al. Myalgic encephalomyelitis/chronic fatigue syndrome patients exhibit altered T cell metabolism and cytokine associations. J Clin Invest. 2020;130(3):1491–505.
11. 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.
12. Karhan E, Gunter CL, Ravanmehr V, Horne M, Kozhaya L, Renzullo S, et al. Perturbation of effector and regulatory T cell subsets in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS). bioRxiv. 2019:2019.12.23.887505.
13. 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.
14. 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.

15. 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.
16. 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.
17. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Network. https://mecfs.rti.org/.
18. mapMECFS. https://www.mapmecfs.org/.
19. CKAN. https://ckan.org/.
20. AWS. https://aws.amazon.com/.
21. Github. https://github.com/search?q=user%3Ackan+is%3Apr+author%3Actownsen357+author%3Ahardingalexh+author%3Ag-rock+author%3Amichael-long88&type=Issues.
22. 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.
23. 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 Res. 2019;48(D1):D440-D4.
24. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005;33(Database issue):D54-8.
25. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform. 2015;7:23.
26. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41(Database issue):D456-63.
27. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007;35(Database issue):D521-6.
28. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155-D62.
29. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.
30. Biomedical Data Repositories and Knowledgebases. https://datascience.nih.gov/data-ecosystem/biomedical-data-repositories-and-knowledgebases.
31. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:1177932219899051.

Figures

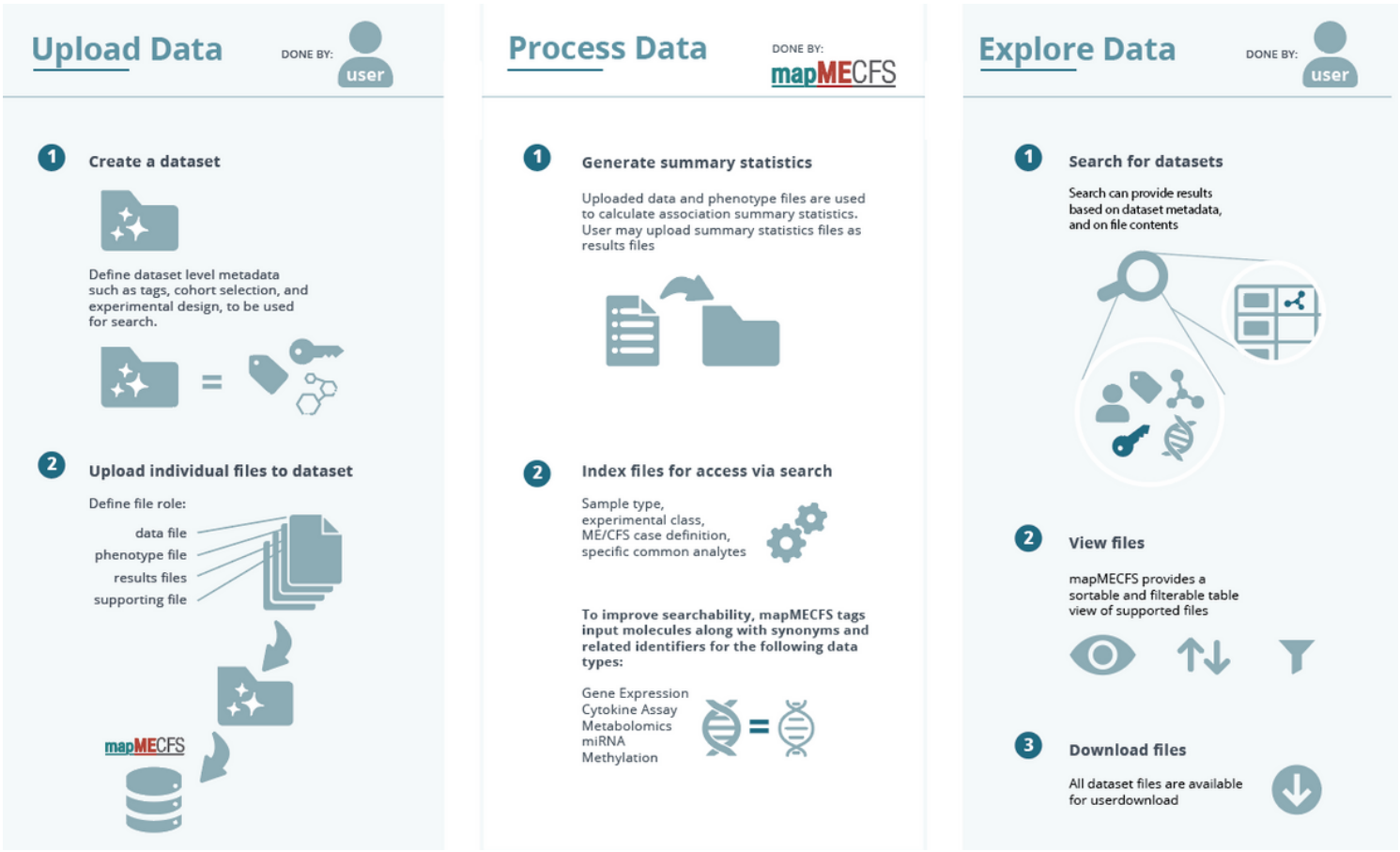

Figure 1

mapMECFS website overview. Data flow chart for mapMECFS. The user uploads files (data, phenotype, results and supporting files) to a Dataset along with metadata. User uploaded files All datafiles are access protected to public or private, depending on the user's preference. mapMECFS processes the data to generate summary statistics and conduct synonym tagging based on the data type. The user can search available data along with viewing, filtering, or downloading files.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- AdditionalFile2.pdf
- AdditionalFile3.png
- AdditionalFile4.png