Use Cases:

We will work with the following use cases, and we will refine and expand them as needed, often based on feedback from C2M2-CC and other DCCs.

Use case 1 (shared with Gene working group)**:** Maurya et al. recently published work on shear stress that causes atherosclerotic plaques in blood vessels where they identified temporally regulated networks that lead to aberrant phenotypes (Maurya et al.[1]). They identified KLF4, a canonical endothelial gene which is downregulated under disturbed flow leading to initiation of oxidative stress, inflammation, altered cell cycle and angiogenesis. Several cell types are involved in the formation of blood vessels and the authors wanted to explore the GTEx resource if the altered genes in this pathway are highly expressed in other cell types (tissues) associated with the blood vessel. Further, the authors had previously identified a lncRNA, called LEENE[2], which was associated with the expression of KLF4 and eNOS. The authors wanted to explore if any exRNAs were a derivative of LEENE or other related RNA species. The authors had also investigated key transcriptional networks[3] and wanted to explore if the key TFs were involved in association with H3K4Me3 or H3K27Me3 marks from ChIP-seq data collected by the ENCODE project. The authors also wanted to identify any druggable targets within the altered genes that they may be able to modulate with small molecules. In order to pose these questions, the key identifiers across the data resources are the genes involved. Hence, it is essential that there is a strategy to identify these genes in the CF data resources (satisfying findability, accessibility, and interoperability). A gene-centric query across the DCCs and external resources will help find the information sought above.

Use case 2a. A biomedical researcher is interested in exploring the adverse effects of a drug to treat breast cancer, by using animal and cell models and wishes to specifically seek if the treatment will affect physical activity of a patient post-treatment. The researcher also wishes to design key perturbation experiments to identify other targetable factors that already have FDA approved candidate molecules[4]. This is not an unusual “Use Case” scenario in biomedical research. There are several elements involved here – identification of relevant publicly available data, ensuring the context through available meta data, collation of all relevant data in a seamless manner (in most cases not involving complex software issues), query and analysis of the data to address the above questions, exploration of FDA approved drugs targeting key molecules identified above, and importantly presentation of the analysis in a manner that will facilitate the next experimental design steps. We present a scenario that will address the above. LINCS has proteomic, transcriptomic and epigenomic data on breast cancer cells treated with drugs, the Human Variome Project data has information on genetic perturbations, TCGA on the somatic mutations, MoTrPAC on data pertaining to physical activity, albeit in normal physiology, ENCODE on key regulatory aspects involving the epigenome, and data resources such as DrugBank provide the catalog of FDA approved drugs, their targets and mechanisms. What would be immensely valuable is an infrastructure that contains, a) a query system for discovering pertinent data, b) a system for virtually obtaining and collating such data, c) a workflow system for carrying out the analysis and generating potential hypotheses, and d) a system for presenting the results in a manner that would allow the researcher to interactively assess the results and develop potential hypotheses.

Use case 2b. The same researcher now uses a breast cancer cell line BT474 to study resistance and sensitivity to the drug Trastuzamab. The researcher finds the cell lines behave differently when treated further with growth factors IGF, HGF and HRG (the known factors involved in tumorigenic transformation and wishes to identify mechanisms associated with sensitivity and resistance. There is relevant data present in LINCS, TCGA (from NCI), IDG, GTex, and 4DN. While there are tools that have the ability to integrate some of these data (such as LJP-BCNB and DEGenR), the are not readily accessible. Our workflow APIs developed in this base module and in the Playbook Partnership proposals will aid in the user identifying protein networks, transcriptional changes and transcription regulatory modules involved in differentiating sensitivity and resistance to the drug.

Use case 3 A patient visits the cardiologist complaining of dizziness, chest pain, and shortness of breath. The cardiologist suspects ventricular tachycardia. It is confirmed by ECG. Since there is a familial history, the cardiologist genotypes the patient and identifies the R92W-TnT mutation[5], a marker associated with cardiac hypertrophy. A colleague collects fibroblasts from the patient and transforms them via induced pluripotent stem cells into cardiomyocytes, followed by transcriptional studies. Down regulation of the SERCA2A gene and

increased ROS production is observed. The cardiologist would like to design experiments to explore, preferably through their browser: i) whether or not their patient's heart is hypertrophied; ii) whether a calcium channel blocker such as nifedipine[6, 7], or a late sodium current blocker such as ranolazine[8, 9], would be a better treatment; and iii) the appropriate personalized drug dose. As above, the principles are similar, namely, discovery of the data sources, collation and integrative analysis, and presentation of results in a biologist-friendly manner. TOPMED, a NHLBI resource in combination with a human variome database, HuBMaP, GTEx, IDG, and SPARC would have the pertinent data for analysis.

Use case 4 (Shared with Clinical Metadata working group) A clinician researcher is reviewing a 52-year-old female Type II diabetes patient who has been treated for 2 years with metformin. She wishes to identify key genes/functions altered in patients with similar demographic and dietary profile (dbGaP, Framingham study/TOPMED data) as well as identify metabolite differences associated with physiological phenotypes in similar patients (NMDR). She also wishes to see consequences for cardiovascular conditions in such patients (Dallas Heart Study, TEDDY data, MoTrPAC). Discoverability of such data will depend critically on availability of appropriate clinical metadata associated with each of the data resources.

Use case 5 (Shared with Clinical Metadata working group) A researcher who works on enchondromatoses in children is interested in identifying using family/trios information genomic variants in these subjects and seek if the variant is associated with Ollier disease associated with skeletal dysplasia. There is WGS data on these trios in Gabriella Miller Kids First resource (dbGaP). There are also gene expression data in GEO on zone specific gene expression patterns in articular cartilage on 4 normal human donors. The researcher would like to seek if the genes associated with variants (including sequences of trios) are expressed differentially in the GEO resource. The study would be meaningful if the associated metadata associated with age, clinical conditions and other related information are available. What metadata would enable the cross-mapping? Our base module and the ClOVoc partnership will use FHIR profiles, KGs and APIs developed here to answer such queries.

New use cases in the cycle 2022-2023:

Use case 6: Using CFDE data to understanding role of PNPLA3 in Nonalcoholic fatty liver disease (NAFLD): Patatin-like phospholipase domain-containing protein 3 (PNPLA3) gene encodes for the protein adiponutrin, found in fat and liver cells. PNPLA3 regulates the development of adipocytes and the production and breakdown of fats (lipogenesis and lipolysis) in Metabolism and Glycerophospholipid biosynthesis. Diseases associated with PNPLA3 include Fatty Liver Disease and Nonalcoholic fatty liver disease (NAFLD)[10-13]. The spectrum of NAFLD includes steatosis, nonalcoholic steatohepatitis (NASH), and cirrhosis. Recognition and timely diagnosis of these different stages, particularly NASH, is important for both potential reversibility and limitation of complications. Towards integrative analysis to identify PNPLA3 associated mechanisms participating in the progression of NAFLD, we believe data from various DCCs of the NIH Common Fund Data Ecosystem (CFDE) and other resources can be leveraged. We will leverage our experience in multiomics analysis on NAFLD data[10]. There, using linear discriminant analysis, we identified a panel of 20 plasma metabolites (glycerophospholipids, sphingolipids, sterols, etc.) that can be used to differentiate between NASH and steatosis. Our proposed approach is multipronged: 1) Identify data on PNPLA3 available at various DCCs (e.g., Metabolomics, GTEx, LINCS, ERCC (for variants information)), 2) starting with PNPLA3, find the interacting proteins using StringDB providing a protein-protein interaction (PPI) network, 3) find the drugs targeting these proteins (LINCS L1000KD2), 4) enrichments against various pathways, functions and diseases, in GTEx (resulting in a ranked list of tissue), and in ChEA and 5) find reactions, enzymes, metabolites and metabolomics studies related to the genes in the PPI network. ChEA will help us identify the transcription factors (TFs) that regulate the genes in the PPI network, from which we can further find other targets enriched in a specific tissue (e.g., liver in this case, using GTEx data). Enrichment of metabolites will help us focus on key metabolic pathways. In principle, some of the above steps can be repeated on combinations of gene lists. Following this comprehensive approach, we hope to gain a mechanistic understanding of how genomic variants in PNPLA3, eventually result in the dysregulated metabolism observed in NAFLD. We also hope to showcase how the data and tools from various CFDE DCCs and other resources may be utilized to study important diseases at a molecular- to functional- to mechanistic level.

Use case 7: Role of PPARG and LXR in diabetes: It is common for a biomedical researcher to investigate a candidate gene(s) implicated in a disease, to construct a disease model, understand its role in the causal mechanism/disease process, and identify therapeutic targets that can ameliorate the condition. This requires identifying and integrating multimodal datasets using facets common to the datasets and relevant to the disease, to validate the hypothesis and enable further discovery, such as identifying biomarkers, therapeutic agents and disease surveillance. The CFDE products (data, tools and services) and the underlying FAIR framework enable such an investigation. For instance, a bioinformatics analysis identified two genes, PPARG (Peroxisome Proliferator-Activated Receptor Gamma) and LXR (Liver X Receptor) as highly enriched and significant in type 2 diabetes (T2D). A potential use case of CFDE resources may have the following components:

- If the researcher wants to construct a cell model system to validate the role of PPARG and LXR, they could use GTEx resources to find the tissues that express these genes in abundance and select the ones that is relevant to the disease and easily accessible (e.g., for PPARG: subcutaneous adipocyte as opposed to omental or mammary adipocyte).
- Further, they could use HuBMAP resources to find adipocyte-specific markers (ex: calbindin 2/calretinin and S100 calcium-binding protein B) to isolate the cell from subcutaneous tissue. Further, they could see the list of drugs affecting the gene from LINCS1000 (FDA-approved drugs/small molecules: both up-regulating and down-regulating the gene), as both PPARG and LXR have been identified as potential drug targets in the treatment of T2D. Drugs that activate PPARG, such as thiazolidinediones, are used as insulin sensitizers. Similarly, LXR agonists improve glucose homeostasis and insulin sensitivity.
- Genetic variants in PPARG and LXR have been associated with an increased risk of T2D. The researchers can find the genetic variants from ClinVar and assess someone's risk of developing T2D.
- Both PPARG and LXR are involved in lipid metabolism, including lipid uptake, fatty acid oxidation, and adipogenesis. Therefore, the researchers could find deranged metabolites in T2D in the metabolomics workbench data and compare that with healthy individuals from MoTrPAC studies. This may help identify metabolic signatures of T2D and biomarkers for early diagnosis and disease surveillance.

Use Case 8: Role of IDH1 in various cancers: The effects of *IDH* [isocitrate dehydrogenase (NADP(+))] mutations not only span epigenetic, differentiation and metabolic programmes, but are highly prevalent across a variety of cancer types of different origins such as low-grade glioma and secondary glioblastoma, acute myeloid leukemia, cholangiocarcinoma, chondrosarcoma, sinonasal undifferentiated carcinoma and angioimmunoblastic T cell lymphoma[14-16]. They are known to occur early during tumorigenesis and are found to have uniform expression in cancer cells, thus making them ideal therapeutic targets. A biomedical researcher interested in investigating the role of IDH in cancer and examining its potential as a therapeutic target would be aided by the CFDE tools and portal, that will enable them to access information, both qualitative and quantitative, under one integrated platform. Here we envision a potential workflow connecting the gene IDH with various CFDE components in the context of cancer. For simplicity, let us consider the cytoplasmic form of IDH, the IDH1 gene, as a use case.

- A possible first step would be to investigate which tissue types express IDH1 and can access expression data in tissue types of interest. This ability is provided by GTEx which enables the user to view tissue and sample statistics pertaining to IDH1.
- Next, the user may be interested in looking at differential gene expression of IDH1 under various knock out conditions (cancer and normal) or potential drug targets associated with IDH1, which is provided by the LINCS resource. It provides differential expression signatures of IDH1 in RNA-seq like experiments and the user can also access statistics regarding drug perturbations (both up and down) associated with the gene.
- Further, given that IDH1 is an isozyme that catalyzes the reversible oxidative decarboxylation of isocitrate to α-ketoglutarate (α-KG) while reducing NADP+ to NADPH, the CFDE resource MetGENE allows the user to access metabolomic studies, obtained from Metabolomics Workbench (MW) pertaining to IDH1 via the pathways and reactions in which IDH1 participates, and the metabolites pertaining to those reactions. One can use the tool MetENP on MW studies to perform statistical analysis (cancer vs. normal) pertaining to the metabolites and carry out further interpretation such as functional enrichment and metabolite-gene association.

- Genetic variants of IDH1 such as those with the mutations R132H, R132C, R132G, R132S and R132L are known to be sensitive to Ivosidenib, which inhibits D-2HG production. These can be studied using ClinVar.

Thus, the CFDE suite of tools will enable biomedical researchers to access various resources under one dashboard enabling them to study the role of IDH1 in various cancers utilizing state-of-the-art omics data and approaches.

Use Case 9: Alteration in energy metabolism in COVID-19 and its impact on diabetes: A recent research article established the correlation between elevated plasma glucose levels favoring SARS-CoV-2 infection and monocyte response via HIF-1a/Glycolysis dependent pathway[17]. A biomedical researcher, using CFDE data, could further verify it by examining the metabolic profile—especially the deranged glucose metabolism in patients with severe COVID-19 infection—using the dataset in Metabolomics Workbench (MW), specifically, data from the project 'PR001469- Integrated metabolic and inflammatory signatures associated with severity, fatality, and recovery of COVID-19'. Similarly, one can examine the similarities or differences in neutrophil metabolism during COVID-19 and compare it with that in monocytes with the study 'PR001600 Human neutrophil metabolomics' (PR001600). Further, the dataset from the project 'PR001474 Serum metabolomics profiling identifies new predictive biomarkers for disease severity in COVID-19 patients' (PR001474). will help develop a predictive model for disease severity given the coexisting condition of diabetes and its control status. Also, the dataset from the study 'ST001921- An Airway Organoid-Based Screen Identifies a Role for the HIF1α-Glycolysis Axis in SARS-CoV-2 Infection' (ST001921) can be used to further validate the role of HIF-1a in COVID-19 pathogenesis. Finally, one could leverage the IDG resource PHAROS to identify HIF-1a targets and develop/validate the hypothesis that anti-H1F-1a would suppress T-cell dysfunction, cytokine storm and monocyte-induced lung epithelial cell death.

Literature Cited

1. Maurya MR, Gupta S, Li JY, Ajami NE, Chen ZB, Shyy JY, Chien S, Subramaniam S. Longitudinal shear stress response in human endothelial cells to atheroprone and atheroprotective conditions. Proc Natl Acad Sci U S A. 2021;118(4). PMCID: 7848718.
2. Miao Y, Ajami NE, Huang TS, Lin FM, Lou CH, Wang YT, Li S, Kang J, Munkacsi H, Maurya MR, Gupta S, Chien S, Subramaniam S, Chen Z. Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. Nat Commun. 2018;9(1):292. PMCID: PMC5773557.
3. Ajami NE, Gupta S, Maurya MR, Nguyen P, Li JY, Shyy JY, Chen Z, Chien S, Subramaniam S. Systems biology analysis of longitudinal functional response of endothelial cells to shear stress. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2017;114(41):10990-5. PMCID: 5642700.
4. Mollah SA, Subramaniam S. Histone Signatures Predict Therapeutic Efficacy in Breast Cancer. IEEE Open J Eng Med Biol. 2020;1:74-82. PMCID: 7207876.
5. Vakrou S, Fukunaga R, Foster DB, Sorensen L, Liu Y, Guan Y, Woldemichael K, Pineda-Reyes R, Liu T, Tardiff JC, Leinwand LA, Tocchetti CG, Abraham TP, O'Rourke B, Aon MA, Abraham MR. Allele-specific differences in transcriptome, miRNome, and mitochondrial function in two hypertrophic cardiomyopathy mouse models. JCI Insight. 2018;3(6). PMCID: 5926940.
6. Luo J, Zhang WD, Du YM. Early administration of nifedipine protects against angiotensin II-induced cardiomyocyte hypertrophy through regulating CaMKII-SERCA2a pathway and apoptosis in rat cardiomyocytes. Cell Biochem Funct. 2016;34(3):181-7.
7. Furberg CD, Psaty BM, Meyer JV. Nifedipine. Dose-related increase in mortality in patients with coronary heart disease. Circulation. 1995;92(5):1326-31.
8. Huang M, Liao Z, Li X, Yang Z, Fan X, Li Y, Zhao Z, Lang S, Cyganek L, Zhou X, Akin I, Borggrefe M, El-Battrawy I. Effects of Antiarrhythmic Drugs on hERG Gating in Human-Induced Pluripotent Stem Cell-Derived Cardiomyocytes From a Patient With Short QT Syndrome Type 1. Front Pharmacol. 2021;12:675003. PMCID: 8138577.

9. Belardinelli L, Shryock JC, Fraser H. Inhibition of the late sodium current as a potential cardioprotective principle: effects of the late sodium current inhibitor ranolazine. Heart. 2006;92 Suppl 4(Suppl 4):iv6-iv14. PMCID: 1861317 of intellectual property rights for ranolazine.
10. Gorden DL, Myers DS, Ivanova PT, Fahy E, Maurya MR, Gupta S, Min J, Spann NJ, McDonald JG, Kelly SL, Duan J, Sullards MC, Leiker TJ, Barkley RM, Quehenberger O, Armando AM, Milne SB, Mathews TP, Armstrong MD, Li C, Melvin WV, Clements RH, Washington MK, Mendonsa AM, Witztum JL, Guan Z, Glass CK, Murphy RC, Dennis EA, Merrill AH, Jr., Russell DW, Subramaniam S, Brown HA. Biomarkers of NAFLD progression: a lipidomics approach to an epidemic. J Lipid Res. 2015;56(3):722-36. PMCID: 4340319.
11. Cohen JC, Horton JD, Hobbs HH. Human fatty liver disease: old questions and new insights. Science. 2011;332(6037):1519-23. PMCID: 3229276.
12. Dong XC. PNPLA3-A Potential Therapeutic Target for Personalized Treatment of Chronic Liver Disease. Front Med (Lausanne). 2019;6:304. PMCID: 6927947.
13. Pingitore P, Romeo S. The role of PNPLA3 in health and disease. Biochim Biophys Acta Mol Cell Biol Lipids. 2019;1864(6):900-6.
14. Calvert AE, Chalastanis A, Wu Y, Hurley LA, Kouri FM, Bi Y, Kachman M, May JL, Bartom E, Hua Y, Mishra RK, Schiltz GE, Dubrovskyi O, Mazar AP, Peter ME, Zheng H, James CD, Burant CF, Chandel NS, Davuluri RV, Horbinski C, Stegh AH. Cancer-Associated IDH1 Promotes Growth and Resistance to Targeted Therapies in the Absence of Mutation. Cell Rep. 2017;19(9):1858-73. PMCID: 5564207.
15. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, Fantin VR, Jang HG, Jin S, Keenan MC, Marks KM, Prins RM, Ward PS, Yen KE, Liau LM, Rabinowitz JD, Cantley LC, Thompson CB, Vander Heiden MG, Su SM. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature. 2009;462(7274):739-44. PMCID: 2818760.
16. Dekker LJM, Wu S, Jurriens C, Mustafa DAN, Grevers F, Burgers PC, Sillevis Smitt PAE, Kros JM, Luider TM. Metabolic changes related to the IDH1 mutation in gliomas preserve TCA-cycle activity: An investigation at the protein level. FASEB J. 2020;34(3):3646-57.
17. Codo AC, Davanzo GG, Monteiro LB, de Souza GF, Muraro SP, Virgilio-da-Silva JV, Prodonoff JS, Carregari VC, de Biagi Junior CAO, Crunfli F, Jimenez Restrepo JL, Vendramini PH, Reis-de-Oliveira G, Bispo Dos Santos K, Toledo-Teixeira DA, Parise PL, Martini MC, Marques RE, Carmo HR, Borin A, Coimbra LD, Boldrini VO, Brunetti NS, Vieira AS, Mansour E, Ulaf RG, Bernardes AF, Nunes TA, Ribeiro LC, Palma AC, Agrela MV, Moretti ML, Sposito AC, Pereira FB, Velloso LA, Vinolo MAR, Damasio A, Proenca-Modena JL, Carvalho RF, Mori MA, Martins-de-Souza D, Nakaya HI, Farias AS, Moraes-Vieira PM. Elevated Glucose Levels Favor SARS-CoV-2 Infection and Monocyte Response through a HIF-1alpha/Glycolysis-Dependent Axis. Cell Metab. 2020;32(3):437-46 e5. PMCID: 7367032.