

Common Fund Data Ecosystem Phase 2 Concept Clearance

Chris Kinsinger

Assistant Director for Catalytic Data Resources

Office of Strategic Coordination



National Institutes of Health

Office of Strategic Coordination–The Common Fund

OSC (Common Fund)

Concept Clearance: Phase 2 Common Fund Program

Title: Common Fund Data Ecosystem (CFDE)

Aims:

1. Enable users to **query across & use** multiple CF data sets
 - a. Data Resource Portal
 - b. Knowledge Portal
 - c. Cloud Workspace
2. **Training and outreach** to bring people to CF data, and train them to work in the cloud through five sub-initiatives:
 - a. Center for Training
 - b. Training Fellowships
 - c. Course Development and Delivery
 - d. Diversity Supplements
 - e. Pilot Projects Enhancing Utility of Common Fund Data
3. **Coordinate and integrate** infrastructure and activities into a cohesive ecosystem
 - a. Integration and Coordination Center

Funds Available: \$23M per year

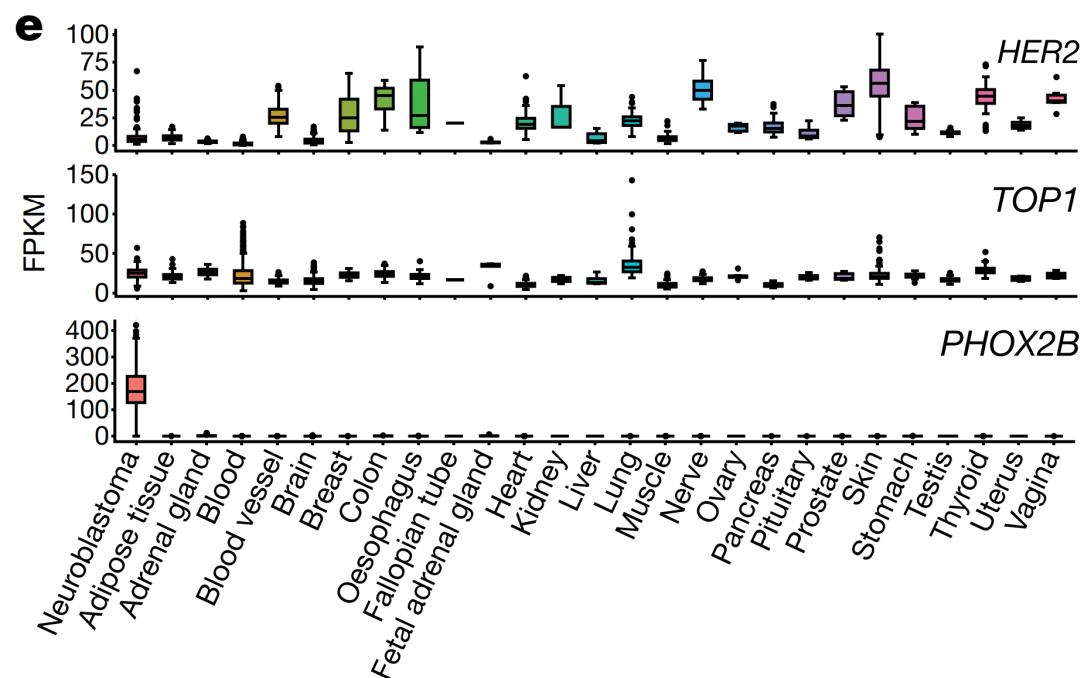
Program Duration: 5 years

Council Action: Vote for approval of the concept for Common Fund Data Ecosystem

CFDE CoC Working Group – Key Observations and Recommendations

- The goals of the CFDE are important to continue.
- CFDE has made substantial progress in its initial phase; it has a solid foundation to continue future work.
- CFDE challenges are common to data science, particularly concerning sustainability of data and resources and inter-operating across platforms
 - Partner with extant repositories for data storage
 - Partner with ODSS, ICs, and other agencies to build toward interoperability
- Transition to cloud computing is important but challenging: Increase and diversify training and outreach initiatives
- The chief metric of success is discovery: If the CFDE is successful, in five years many investigators will be using Common Fund data for new discoveries and new purposes.

Drug targets for neuroblastoma



- Can I find gene expression of neuroblastoma cohorts?



- Can I conduct differential expression analysis with normal tissue?



[Yarmarkovich et al., 2021, Nature 599, 477](#)

Challenges of Data Wrangling



Find and access the data



Harmonize files into one data set



Create and analyze cross-cohort tables

How to find and access the data?

- [CFDE Search Portal](#)
- CAVATICA cloud workspace



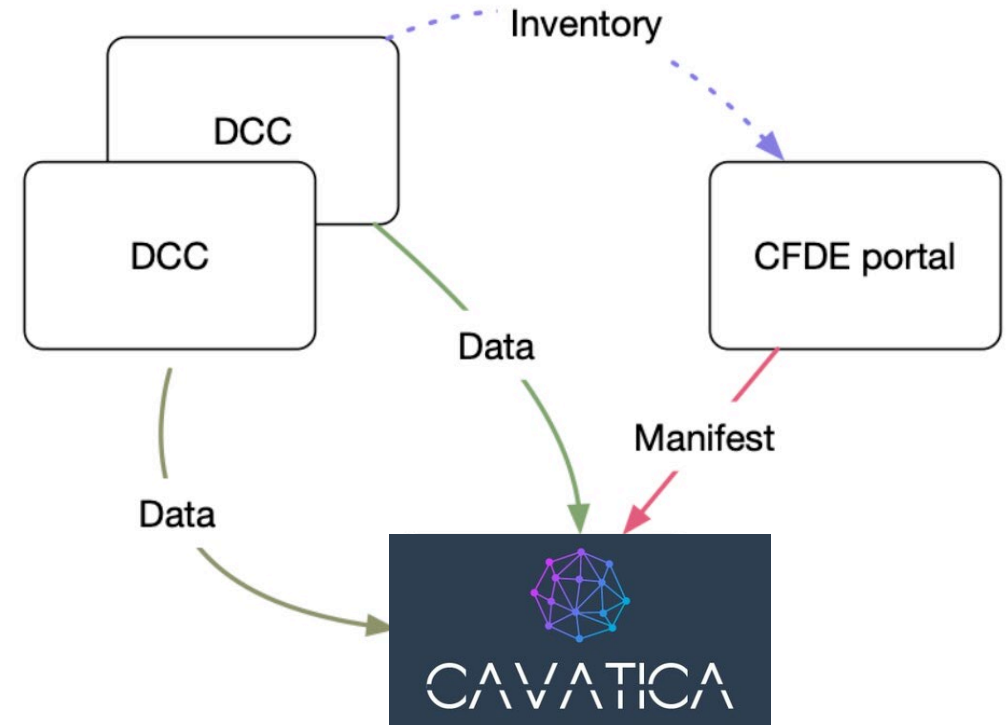
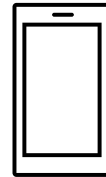
Common Fund Data Ecosystem Search Portal

Find files, biosamples, or subjects from Common Fund data sets



Harmonization and analysis

- Establish common pipelines (e.g., RNA seq)
- Appyters for creating and analyzing standard tables
- [More details are available here](#)



Accelerating and democratizing discovery



~18 months

296 (0.48%) genes with minimum
 $\text{LogFC} > 1$ and $p < 0.01$



33 (11%) genes predicted
to be membrane associated



6 genes with high absolute RNA expression
(FPKM > 50)

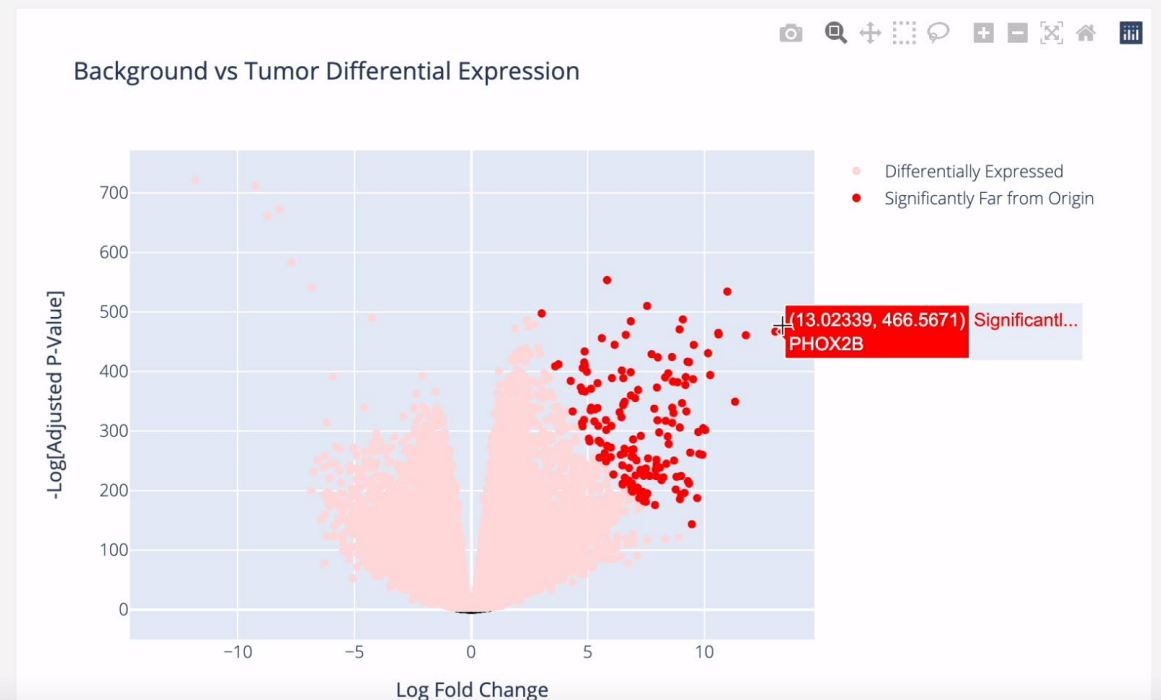
PHOX2B, TH, IGFBPL1, CHRNA3, HMX1,
GFRA2



~3 hours

Narrow Down Candidate Set

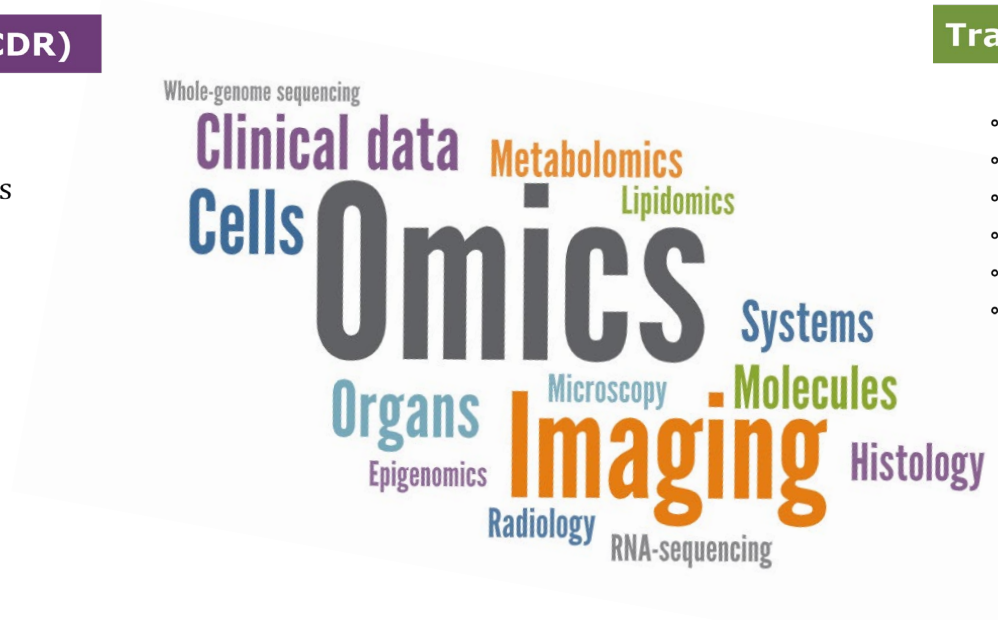
We identify significantly differentially expressed genes with $\text{logFC} \cdot t\text{-statistic}$ products which significant deviate from the mean, or equivalently, those points which are furthest from the volcano plot origin.



CF Programs Generate a Wide Array of Data Types

Catalytic Data Resources (CDR)

- Bridge2AI
- CFDE
- HuBMAP
- Kids First
- Metabolomics



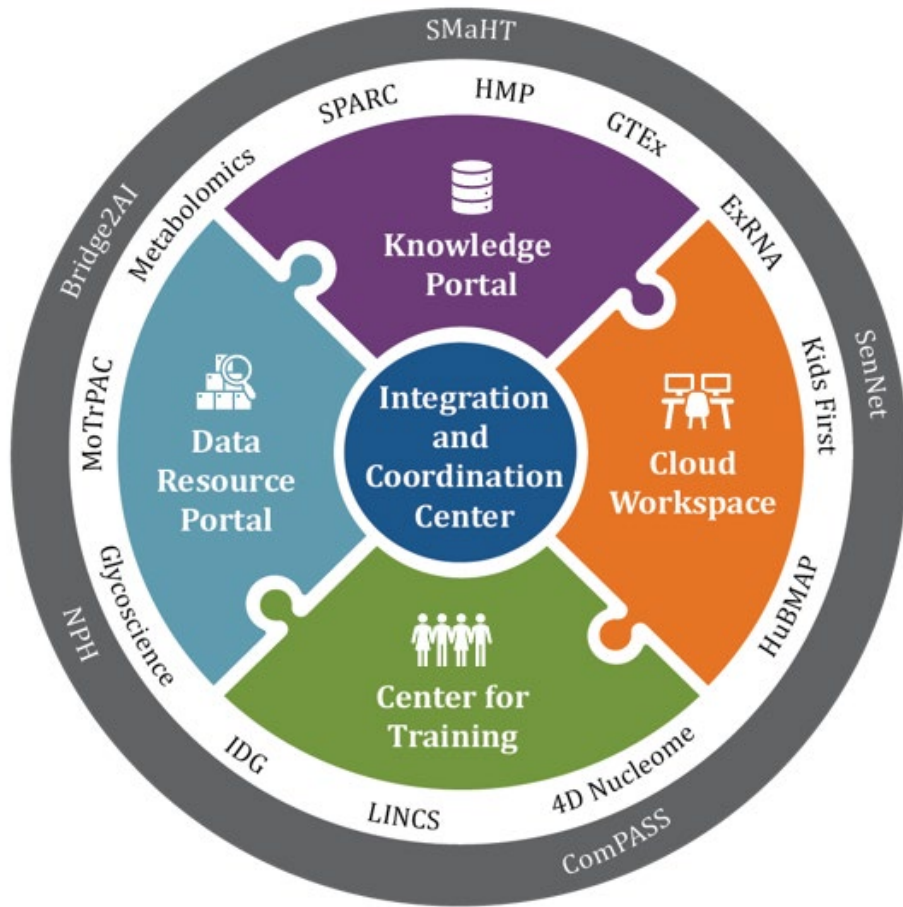
Transformational Science and Discovery (TSD)

- 4D Nucleome
- A2CPS
- CryoEM
- DS-I Africa
- ExRNA
- Global Health
- Global Health
- Glycoscience
- HRHR
- IDG
- KOMP2
- MoTrPAC
- NPH
- SMaHT
- SPARC

Re-Engineering the Research Enterprise (RRE)

- DPC
- FIRST
- SCGE
- Transformative Research to Address Health Disparities and Advance Health Equity
- UDN

Proposed Structure of CFDE in Phase II



- 5 interconnecting components
- Continued engagement with DCCs
- Enhancing findability and accessibility of data
- Integration and Coordination Center ensures cohesion and collaboration
- Increased emphasis on training and outreach
 - Center for Training will coordinate with awardees supported by 4 additional Training/Outreach initiatives, bringing them into the ecosystem

Portals for Data and Knowledge Resources

- Create two portals for users to **query across CF programs**
 - Data: Data files and tools
 - Knowledge: Integrated information and knowledge
- Interconnected portals
- Cohesive interface
- Potential to merge into a single portal as they mature

Success Metric: Robust and growing user community for the search and knowledge portal(s), with enhanced utility over time

Cloud Workspace

- Enable users to import their own data to co-analyze with CF data and resources
- Leverage existing solutions
- Serve both novice and expert users
- Perform mid-cycle review

Success Metric: Robust and growing user community, growth in the tools and capabilities to facilitate analysis

Training and Outreach

5 sub-initiatives that will grow and diversify the CF data user-base

1. **Center for Training (Training)**: Scalable course work to address unmet needs, coordinate across all training sub-initiatives
2. **Training Fellowships (Training)**: Two-year diversity fellowships aimed at supporting trainees interested in reusing CF data (4x)
3. **Course Development and Delivery (Training)**: Design and implement 2-3 courses with rigorous evaluation; Target diverse early career investigators to enable them to build research projects around cloud computing using CF data resources (2x)
4. **Diversity Supplements (Outreach)**: Partner with ICOs to support reuse of CF data (5x)
5. **Pilot Projects Enhancing Utility of Common Fund Data (Outreach)**: Promote the reuse of datasets and resources from multiple CF programs (5x)

Success Metric: Uptake of trainings, growth in the number of investigators using CF data and resources

Integration & Coordination Center

The ICC will focus on ensuring internal cohesion within the program and implementing a structured evaluation process to ensure a continuous cycle of improvement

The Integration & Coordination Center will have three major responsibilities:

1. Integration & Coordination
2. Sustainability Services
3. Evaluation

Success Metric: Cohesive program with an annual evaluation process to drive continuous improvement

Budget Request for Phase II

(numbers in 000,000)

Initiative	FY23	FY24	FY25	FY26	FY27	Total
Data Portal	\$1.75	\$1.75	\$1.75	\$1.5	\$1.25	\$8
Knowledge Portal	\$1.25	\$1.5	\$1.75	\$2	\$2.25	\$8.75
Workspace	\$0.75	\$1.5	\$1.5	\$1.5	\$1.5	\$6.75
Training & Outreach	\$7	\$7.9	\$7.7	\$7.7	\$7.1	\$37.4
Integration	\$1.5	\$2.25	\$2	\$2	\$2	\$9.75
DCC Engagement	\$8	\$8	\$8	\$8	\$8	\$40
RMS	\$0.2	\$0.2	\$0.2	\$0.2	\$0.2	\$1
Total	\$20.45	\$23.1	\$22.9	\$22.9	\$22.3	\$111.65

Council Action: Vote for approval of the concept renewal for Common Fund Data Ecosystem

 commonfund.nih.gov

 [@NIHCommonFund](https://www.facebook.com/NIHCommonFund)

 [@NIH_CommonFund](https://twitter.com/NIH_CommonFund)



National Institutes of Health

Office of Strategic Coordination–The Common Fund