

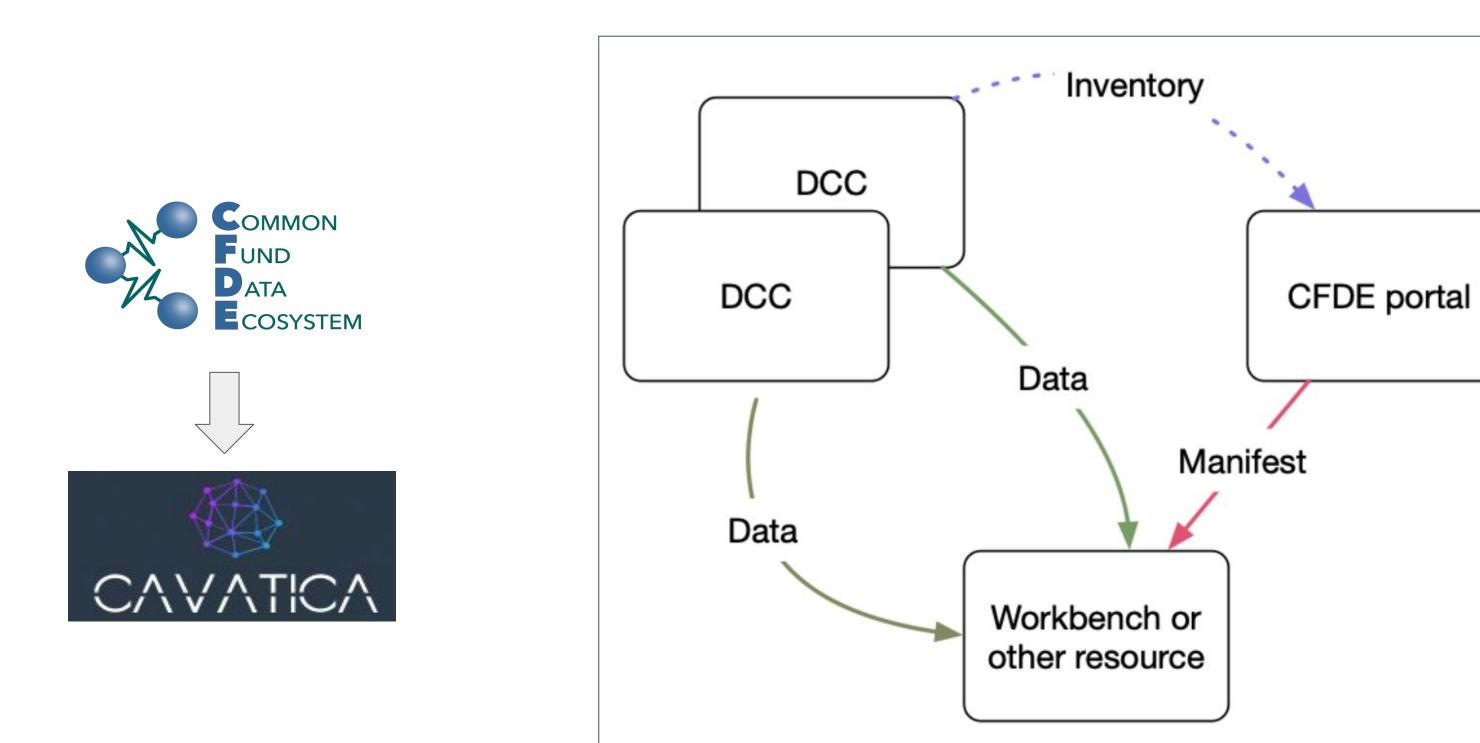
Manifest-based DRS import: A practical solution for cross-DCC dataset analysis to empower translational discovery using Kids First and GTEx data

Surya Saha¹, Michele Mattioni¹, Milos Trboljevac¹, Eric Wenger², Allison Heath², Yuankun Zhu², Robert Carter³, Michelle Giglio³, Suvarna Nadendla³, C. Titus Brown⁴, Bailey K Farrow², Daniel J. B. Clarke⁵, Adam Kraya², Kristin Ardlie⁶, Jared Nedzel⁶, Lan Nguyen⁵, Avi Ma'ayan⁴, Owen White³, Jack DiGiovanna¹, Adam Resnick² Seven Bridges Inc., Charlestown, MA 02129, USA; Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, PA 19146, USA; Institute for Genome Sciences, University of Maryland, Baltimore, MD 21201, USA; School of Veterinary Medicine, University of California, Davis, CA 95616, USA; The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; Department of Pharmacological Sciences, Mount Sinai, New York, NY 10029, USA

Abstract

A key challenge in data discovery is the coordination and assembly of datasets from across Common Fund Data Ecosystem (CFDE) Data Coordination Centers (DCC) in an easy to use and meaningful manner to accelerate usage by researchers. We have implemented a manifest-based import on our CAVATICA platform for a user to create a cross-Common Fund dataset cohort and combine the results with their own data in order to accelerate platform-based discovery and clinical translation. We propose the required field: drs_uri followed by these optional fields: study_registration, study_id, participant_id, specimen_id, experimental_strategy, file_format and fhir_document_reference. The study_registration is the external source of the study_id (e.g. dbGaP). The study_id, participant_id and specimen_id fields are unique identifiers that can be used to retrieve more information. The experimental_strategy and file_format fields are based on the Genomics Data Commons definitions. The fhir_document_reference points to the FHIR Document Reference, if metadata is available on a FHIR server. This process provides an efficient method to import a list of DRS URIs along with relevant metadata. In this use case, a manifest is created from the Common Fund Data Ecosystem portal with GTEx and Kids First (KF) neuroblastoma RNA sequencing assays and brought into a collaborative CAVATICA workspace. The data authorization aspect is managed by CAVATICA. For KF and GTEx datasets which have controlled access, the user's dbGaP access authorizations are checked and the data becomes accessible only if the user has proper authorization. Authorized users can choose to run their own pipelines or use a KF standard pipeline to harmonize and analyze the combined data set. This use case demonstrates how a user can easily search for and generate a cohort across a federated DCC resource framework followed by DRS-based import into CAVATICA collaborative workspace for democratized access and translational knowledge mining.

Cross-DCC dataset analysis

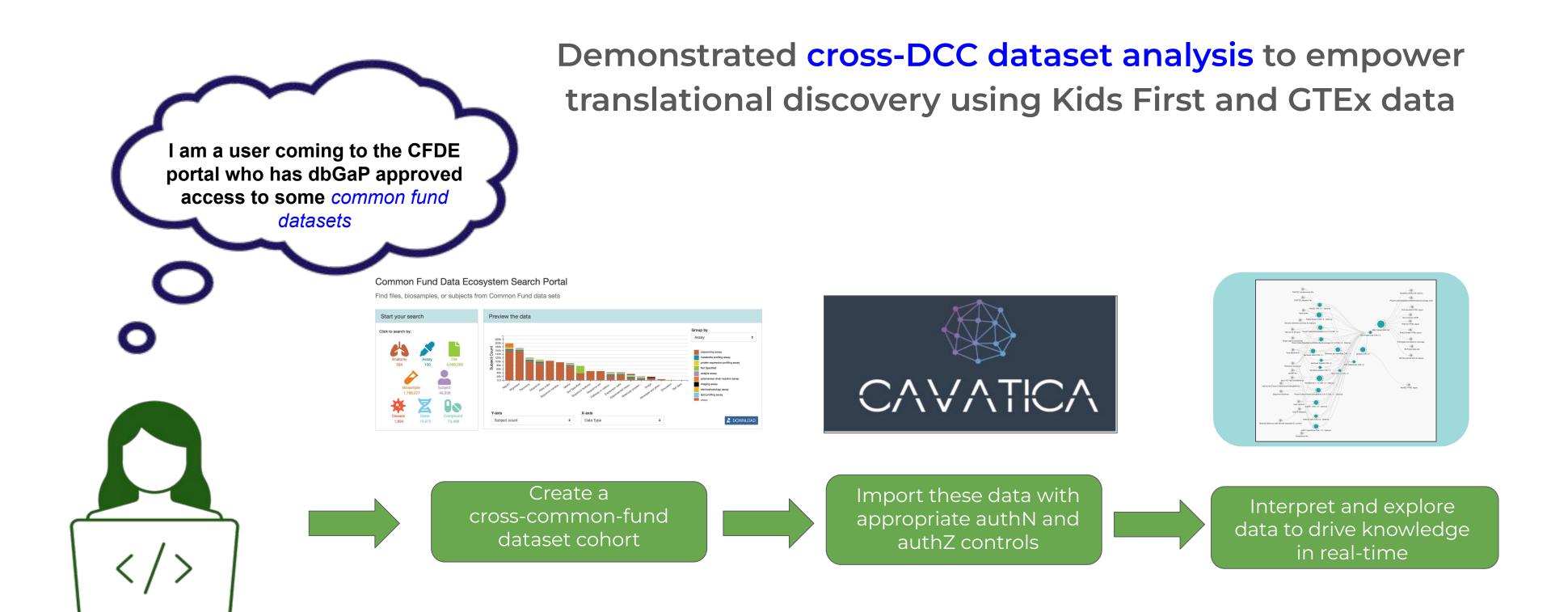




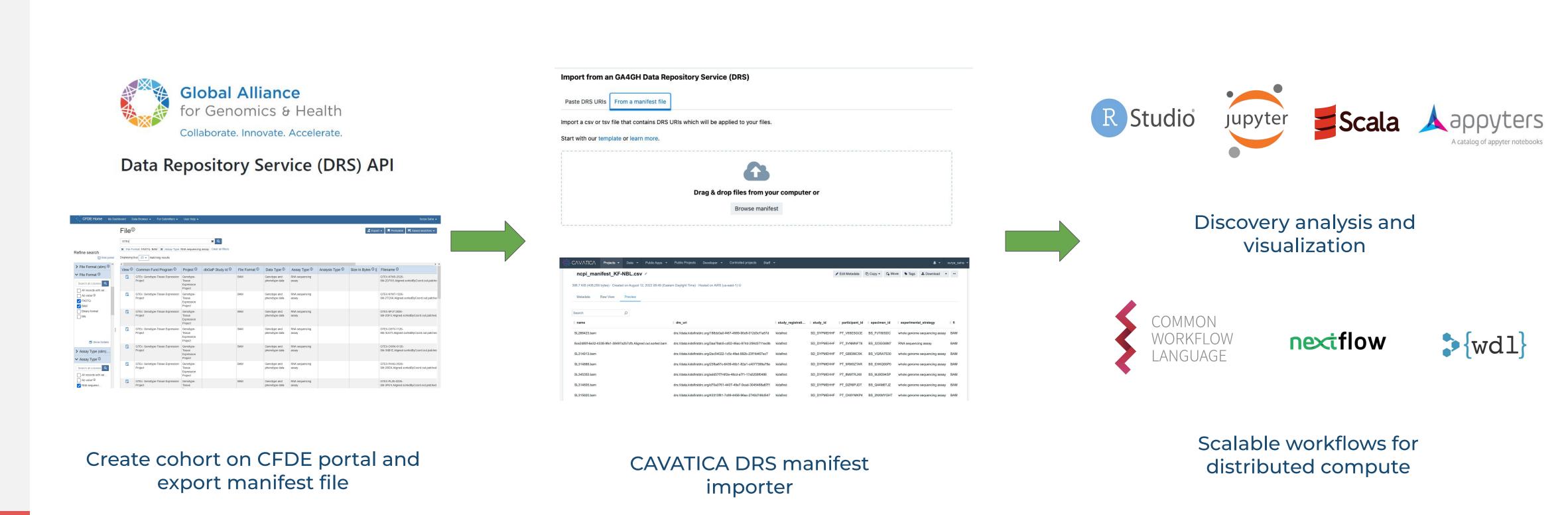








Platform-based discovery powered by DRS



Manifest-based DRS import

Attribute	Definition	Example
drs_uri (required)	DRS URI as defined by GA4GH DRS spec for pointers to file objects.	drs://dg.ANV0:dg.ANV0/97e349fa-88f5-4fef-8008-87 a3aad3560a
study_registration	External source from which the identifier included in study_id originates (answer can be dbGaP for example)	gtex_id
study_id	Unique identifier that can be used to retrieve more information for a study	GTEx
participant_id	Unique identifier that can be used to retrieve more information for a participant	GTEX-N7MS
specimen_id	Unique identifier that can be used to retrieve more information for a specimen	GTEX-N7MS-2526-SM-2D7W3
experimental_strategy	The experimental strategy used to generate the data file referred to by the ga4gh_drs_uri. (Based on GDC definition)	RNA sequencing assay
file_format	The format of the data, see possible values from the data_format fields in GDC. Can use whatever values make sense for the particular implementation.	BAM
fhir_document_reference	Optional fhir url pointing to the FHIR Document Reference, if metadata available on a FHIR Server	https://include-api-fhir-service.includedcc.org/DocumentReference?identifier=HTP.2885a2e1-001b-4751-b622-fdfda2795ea6.rsem.isoforms.results.gz

Next Steps

- Establish dedicated Git repo on official NCPI Github organization
- Socialize the standards to the community to get feedback for further improvements and broader adoption



urya.saha@sevenbridges.co