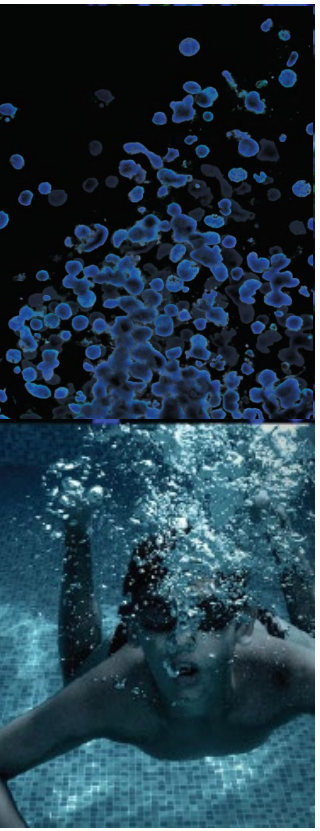
Navigating the HCMI Searchable Catalog and Accessing HCMI Data at NCI's Genomic Data Commons

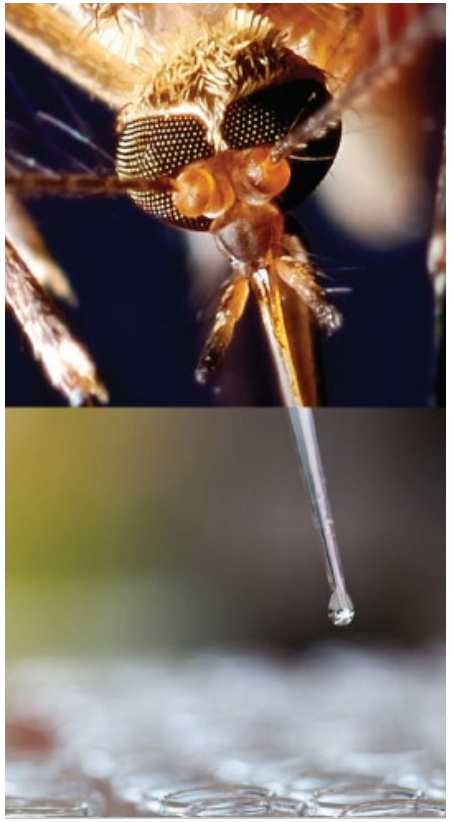
Eva Tonsing-Carter, PhD
Scientific Program Manager
NCI Center for Cancer Genomics

Bill Wysocki, PhD
Lead for User and Data Services
University of Chicago

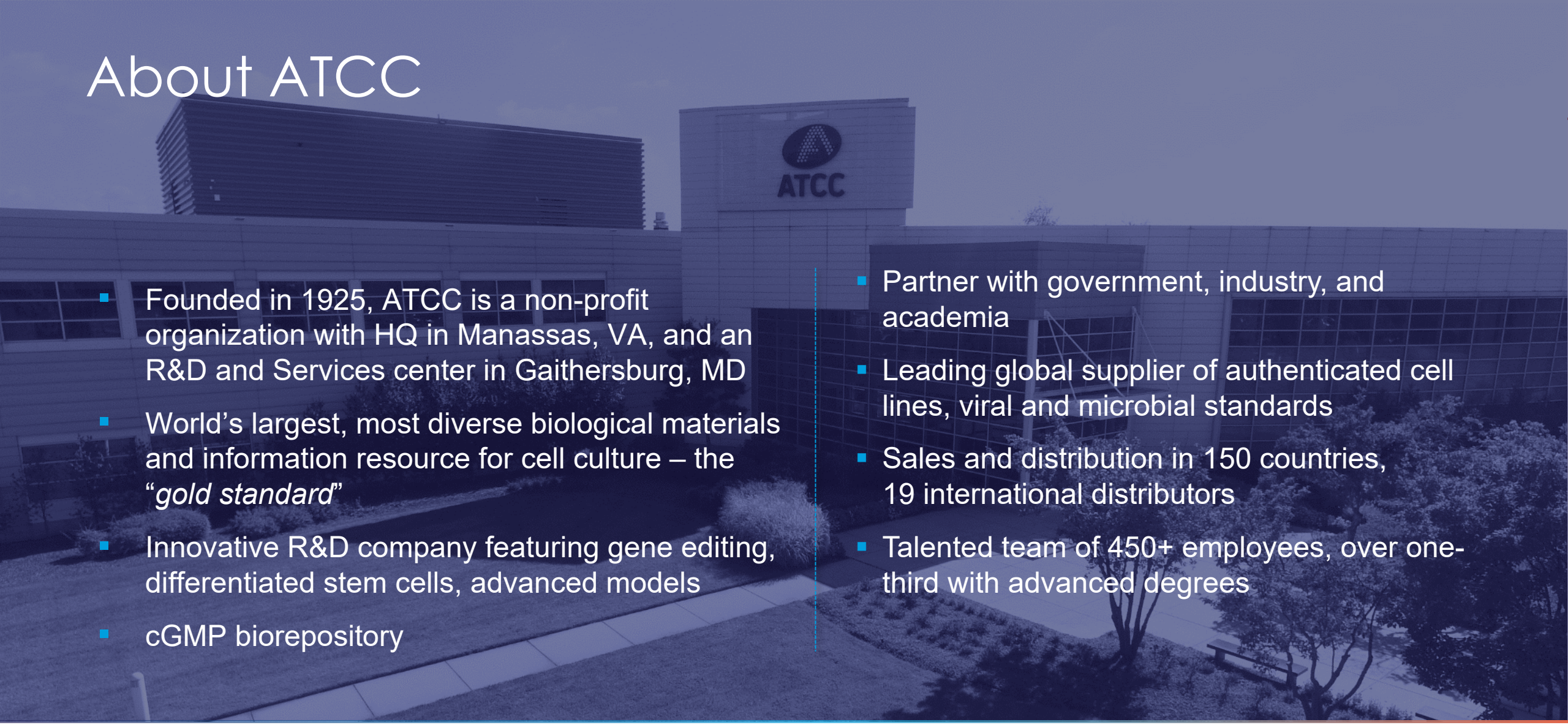
James Clinton, PhD
Lead Scientist
ATCC

Credible Leads to Incredible™





About ATCC



- Founded in 1925, ATCC is a non-profit organization with HQ in Manassas, VA, and an R&D and Services center in Gaithersburg, MD
- World's largest, most diverse biological materials and information resource for cell culture – the "*gold standard*"
- Innovative R&D company featuring gene editing, differentiated stem cells, advanced models
- cGMP biorepository

- Partner with government, industry, and academia
- Leading global supplier of authenticated cell lines, viral and microbial standards
- Sales and distribution in 150 countries, 19 international distributors
- Talented team of 450+ employees, over one-third with advanced degrees

ATCC

Our speakers

Eva Tonsing-Carter, PhD
Scientific Program Manager
National Cancer Institute's Center for Cancer Genomics

NIH NATIONAL CANCER INSTITUTE
Center for Cancer Genomics

Bill Wysocki, PhD
Lead for User and Data Services
Genomic Data Commons in the Center for Translational Data Science (University of Chicago)

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

ATCC

Human Cancer Models Initiative (HCMI): Navigating the HCMI Searchable Catalog and Accessing HCMI Data at NCI's Genomic Data Commons

Eva Tonsing-Carter, Ph.D.

Health Scientist Administrator (Program Officer)

Center for Cancer Genomics

National Cancer Institute

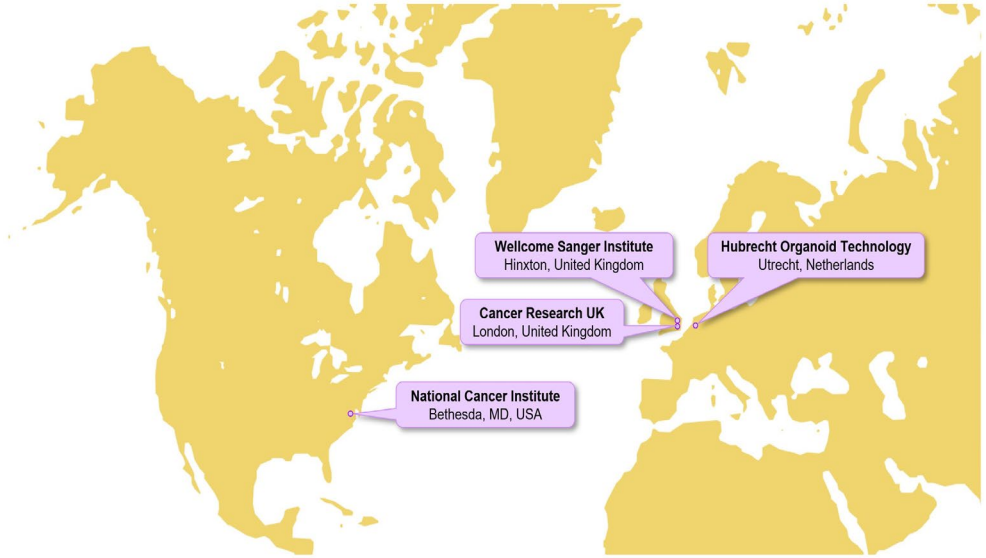
NIH NATIONAL CANCER INSTITUTE

August 21, 2023

Human Cancer Models Initiative (HCMI) Program

- **International consortium** established by:
  - National Cancer Institute (NCI)
  - Cancer Research UK (CRUK)
  - Wellcome Sanger Institute (WSI)
  - Hubrecht Organoid Technology (HUB)

- The consortium's **objective** is to provide next-generation cancer models derived from patient tissues as a scientific community resource.
  - provide models derived from a range of tumor types
  - models are associated with clinical and molecular characterization data
  - models along with associated data will be broadly available to the scientific community



NCI-supported Cancer Model Development Pipeline

Tissues

- Model development
- Nucleic acid isolation

+

Clinical Data

- Controlled vocabulary Case Report Forms (CRFs)
  - Enrollment
  - Follow-up (at least 3 months)
- Quality checked
- Mapped to GDC dictionary

+

Molecular Characterization

- RNA-seq: parent tumor and derived model
- 15X-30X WGS: normal tissue, parent tumor, and derived model
- 150X WXS: normal tissue, parent tumor, and derived model
- DNA methylation array: subset of derived models and parent tumor

+

Data Storage and Harmonization

- Data are harmonized and analyzed through GDC's analytical pipeline
- All available case-associated data are stored at GDC (clinical, biospecimen, and molecular characterization)

=

HCMI Next-Generation Cancer Models as a Community Resource

- Distributed through ATCC
- HCMI Searchable Catalog provides a database of available models

https://www.cancer.gov/ccg/research/functional-genomics/hcmi/about/cancer-model-development

HCMI Resources

- HCMI Searchable Catalog: https://hcmi-searchable-catalog.nci.nih.gov
- Office of Cancer Genomics HCMI Resources: https://www.cancer.gov/ccg/research/functional-genomics/hcmi/using-hcmi
  - Cancer-specific Case Report Forms
  - Informed Consent Template
  - HCMI Searchable Catalog User Guide
- Model Distributor: www.atcc.org/hcmi
  - HCMI models, protocols, and standardized reagents
- Genomic Data Commons (GDC): https://portal.gdc.cancer.gov/projects/HCMI-CMDC
  - Clinical, biospecimen, and molecular characterization data

HCMI Resources – Case Report Forms

Case Report Forms

Human Cancer Models Initiative's (HCMI) cancer type specific case report forms (CRFs) have been developed through collaborations with international clinical experts and the clinical data elements have been standardized through the Cancer Data Standards Registry and Repository (caDSR). As the tumor types modeled through the HCMI are constantly updated, be sure to check back regularly for additional or updated CRFs.

Please note that in some cases, it is possible to collect tissues from different tumor sites (e.g., primary, metastatic, or recurrent) from the same patient for model development. New CRFs capture information about multiple models developed from the same patient within a single form. The multiple model CRFs are designated with "-multi" within the file name. Versioning is used if there are any subsequent edits to the CRFs.

The cancer types that are not highly represented in the HCMI models are considered "rare" by HCMI, however, they may not be considered "rare" based on their representation among cancer incidence rates. The "Rare Cancer CRFs" are being used to capture clinical data for these cancer types: Bladder Cancer, Chordoma, Epithelioid Sarcoma, Extrahepatic Cholangiocarcinoma, Gallbladder Cancer, Spindle Cell Sarcoma, Small Intestine Cancer, and Thyroid Cancer.

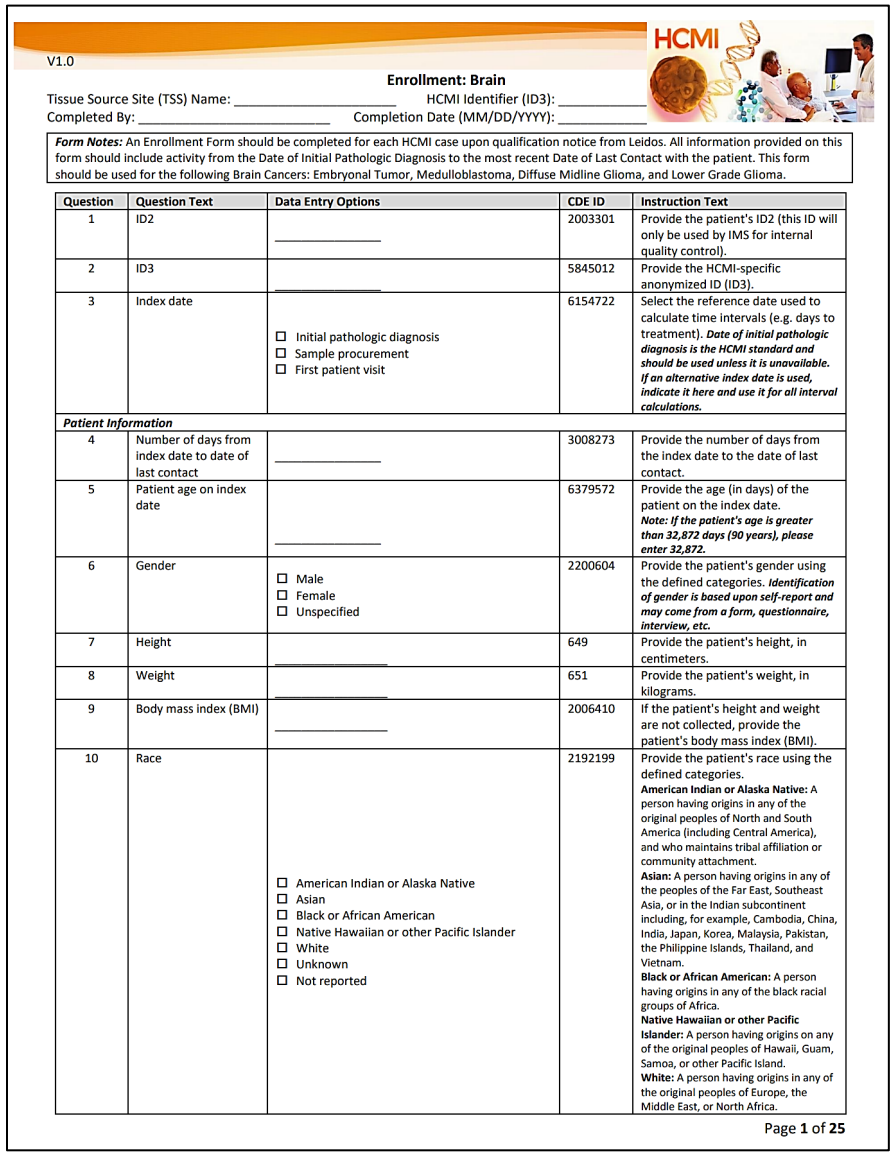
Enrollment Forms

The enrollment CRFs data are collected at the time of patient index date, which includes initial pathologic diagnosis, sample procurement, or first patient visit.

- Ampulla of Vater-multi Enrollment CRF
- Brain-multi Enrollment CRF
- Breast-multi Enrollment CRF
- Colorectal-multi Enrollment CRF
- Endometrium-multi Enrollment CRF
- Esophageal-multi Enrollment CRF
- Ewing Enrollment CRF
- Glioblastoma-multi Enrollment CRF
- Head and Neck-multi Enrollment CRF
- HCC and Intrahepatic-multi Enrollment CRF
- Kidney-multi Enrollment CRF
- Lung-multi Enrollment CRF
- Melanoma-multi Enrollment CRF
- Mesothelioma-multi Enrollment CRF
- Neuroblastoma Enrollment CRF
- Neuroendocrine Enrollment Suppl CRF
- Osteosarcoma Enrollment CRF
- Ovarian-multi Enrollment CRF
- Pancreas-multi Enrollment CRF
- Pediatric Liver Enrollment CRF
- Prostate-multi Enrollment CRF
- Rare-multi Enrollment CRF
- Rhabdomyosarcoma-multi Enrollment CRF
- Stomach-multi Enrollment CRF
- Wilms Enrollment CRF

Follow-up Forms

The follow-up CRF data are collected at least 3 months after enrollment data collection following model establishment.



V1.0

HCMI

Enrollment: Brain

Tissue Source Site (TSS) Name: _____ HCMI Identifier (ID3): _____

Completed By: _____ Completion Date (MM/DD/YYYY): _____

Form Notes: An Enrollment Form should be completed for each HCMI case upon qualification notice from Leidos. All information provided on this form should include activity from the Date of Initial Pathologic Diagnosis to the most recent Date of Last Contact with the patient. This form should be used for the following Brain Cancers: Embryonal Tumor, Medulloblastoma, Diffuse Midline Glioma, and Lower Grade Glioma.

Question	Question Text	Data Entry Options	CDE ID	Instruction Text
1	ID2	_____	2003301	Provide the patient's ID2 (this ID will only be used by IMS for internal quality control).
2	ID3	_____	5845012	Provide the HCMI-specific anonymized ID (ID3).
3	Index date	☐ Initial pathologic diagnosis ☐ Sample procurement ☐ First patient visit	6154722	Select the reference date used to calculate time intervals (e.g. days to treatment). ***Date of initial pathologic diagnosis is the HCMI standard and should be used unless it is unavailable. If an alternative index date is used, indicate it here and use it for all interval calculations.***
Patient Information				
4	Number of days from index date to date of last contact	_____	3008273	Provide the number of days from the index date to the date of last contact.
5	Patient age on index date	_____	6379572	Provide the age (in days) of the patient on the index date. ***Note: If the patient's age is greater than 32,872 days (90 years), please enter 32,872.***
6	Gender	☐ Male ☐ Female ☐ Unspecified	2200604	Provide the patient's gender using the defined categories. ***Identification of gender is based upon self-report and may come from a form, questionnaire, interview, etc.***
7	Height	_____	649	Provide the patient's height, in centimeters.
8	Weight	_____	651	Provide the patient's weight, in kilograms.
9	Body mass index (BMI)	_____	2006410	If the patient's height and weight are not collected, provide the patient's body mass index (BMI).
10	Race	☐ American Indian or Alaska Native ☐ Asian ☐ Black or African American ☐ Native Hawaiian or other Pacific Islander ☐ White ☐ Unknown ☐ Not reported	2192199	Provide the patient's race using the defined categories. **American Indian or Alaska Native:** A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment. **Asian:** A person having origins in any of the peoples of the Far East, Southeast Asia, or in the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. **Black or African American:** A person having origins in any of the black racial groups of Africa. **Native Hawaiian or other Pacific Islander:** A person having origins on any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Island. **White:** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

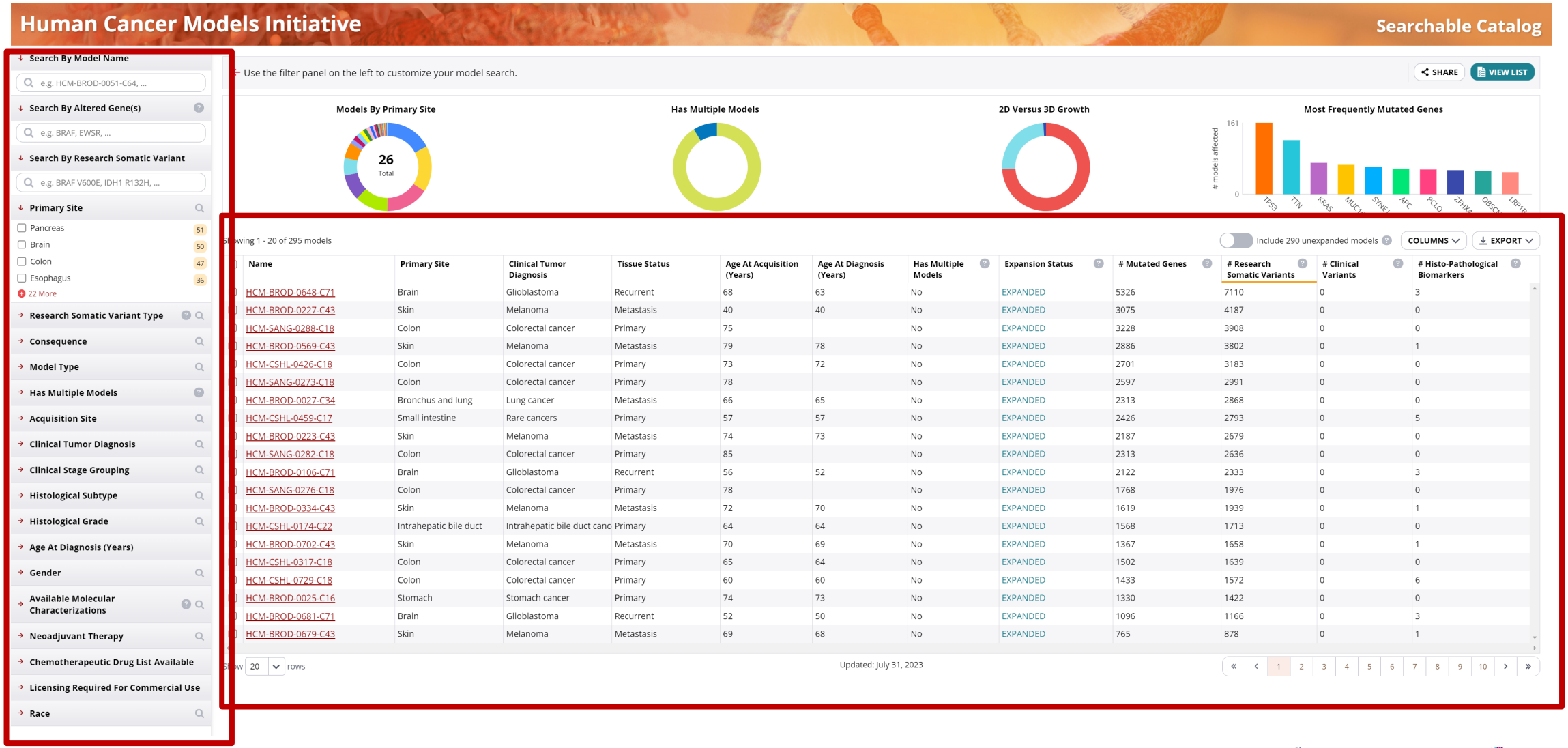
Page **1** of **25**

https://www.cancer.gov/ccg/research/functional-genomics/hcmi/using-hcmi

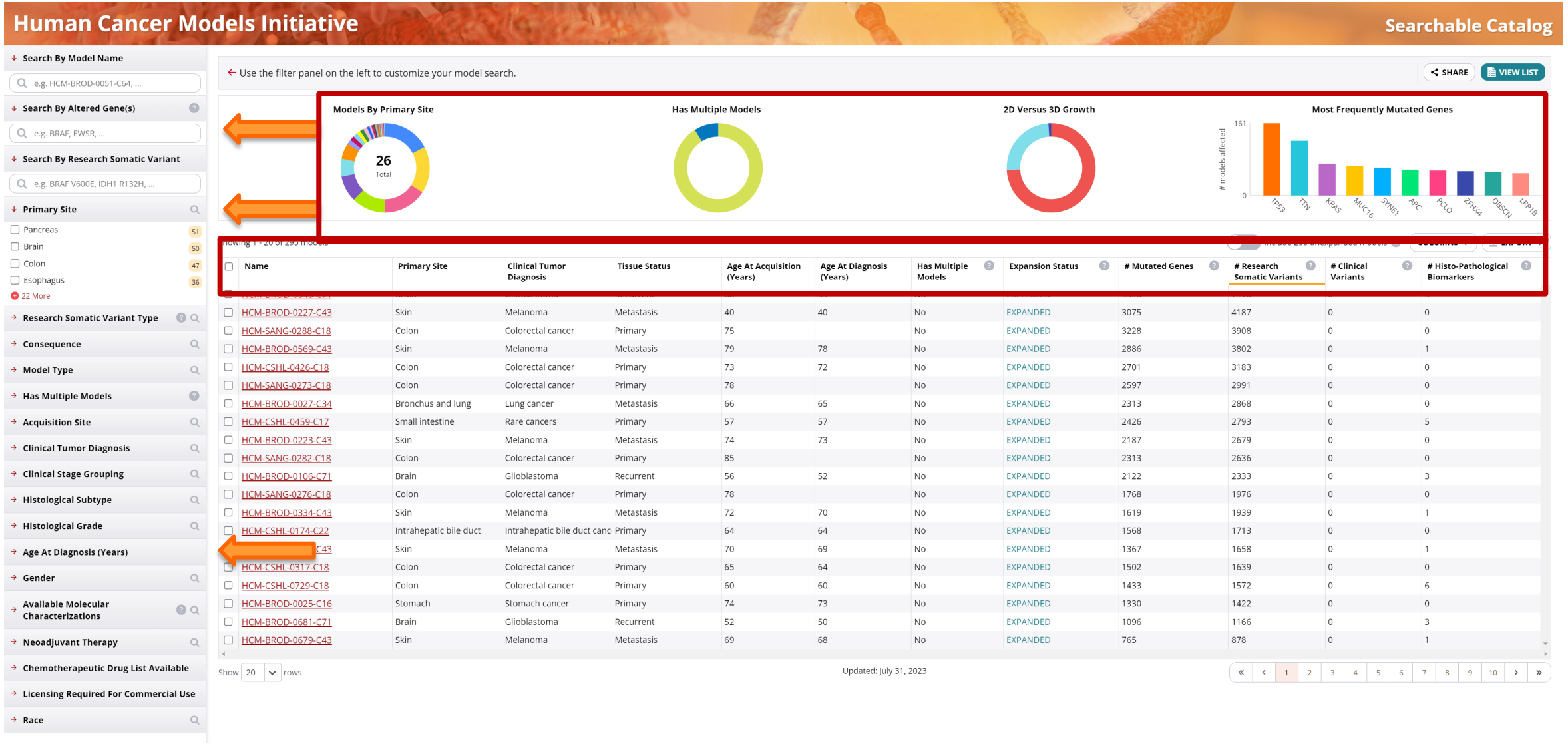
HCMI Searchable Catalog Resource

- **Objective: develop a public-facing online catalog to query and identify HCMI next-generation cancer models available to the research community**
  - The HCMI Searchable Catalog contains model information (e.g. model type), clinical data (e.g. clinical diagnosis, primary site, etc.), distributor information (e.g. images), and molecular characterization data information.
- Developed by OICR (NCI-sponsored FNLCR managed subcontractor)
- Work in progress and will contain model data from all consortium members
- HCMI Searchable Catalog: https://hcmi-searchable-catalog.nci.nih.gov

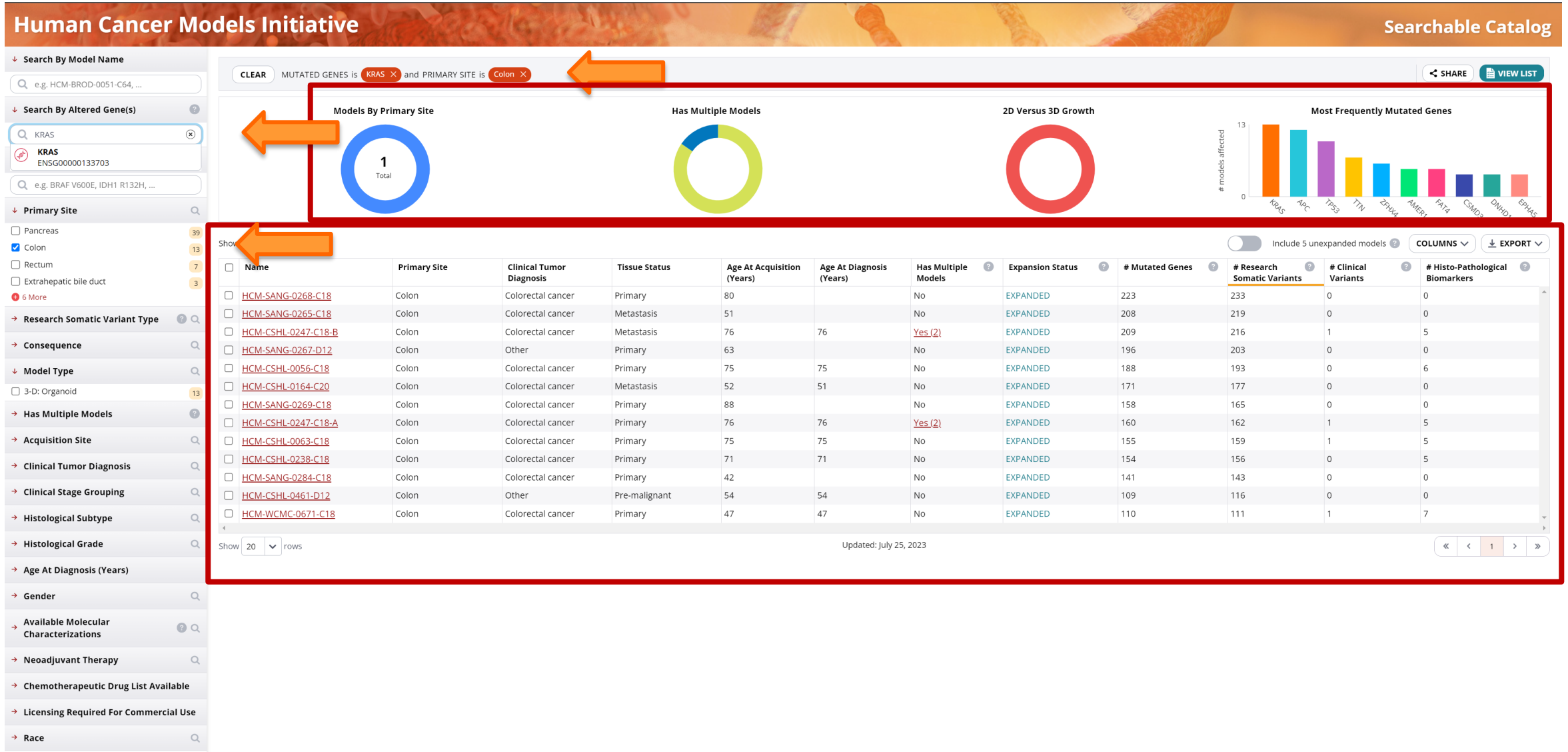
Landing Page: Includes All Available Models to Query



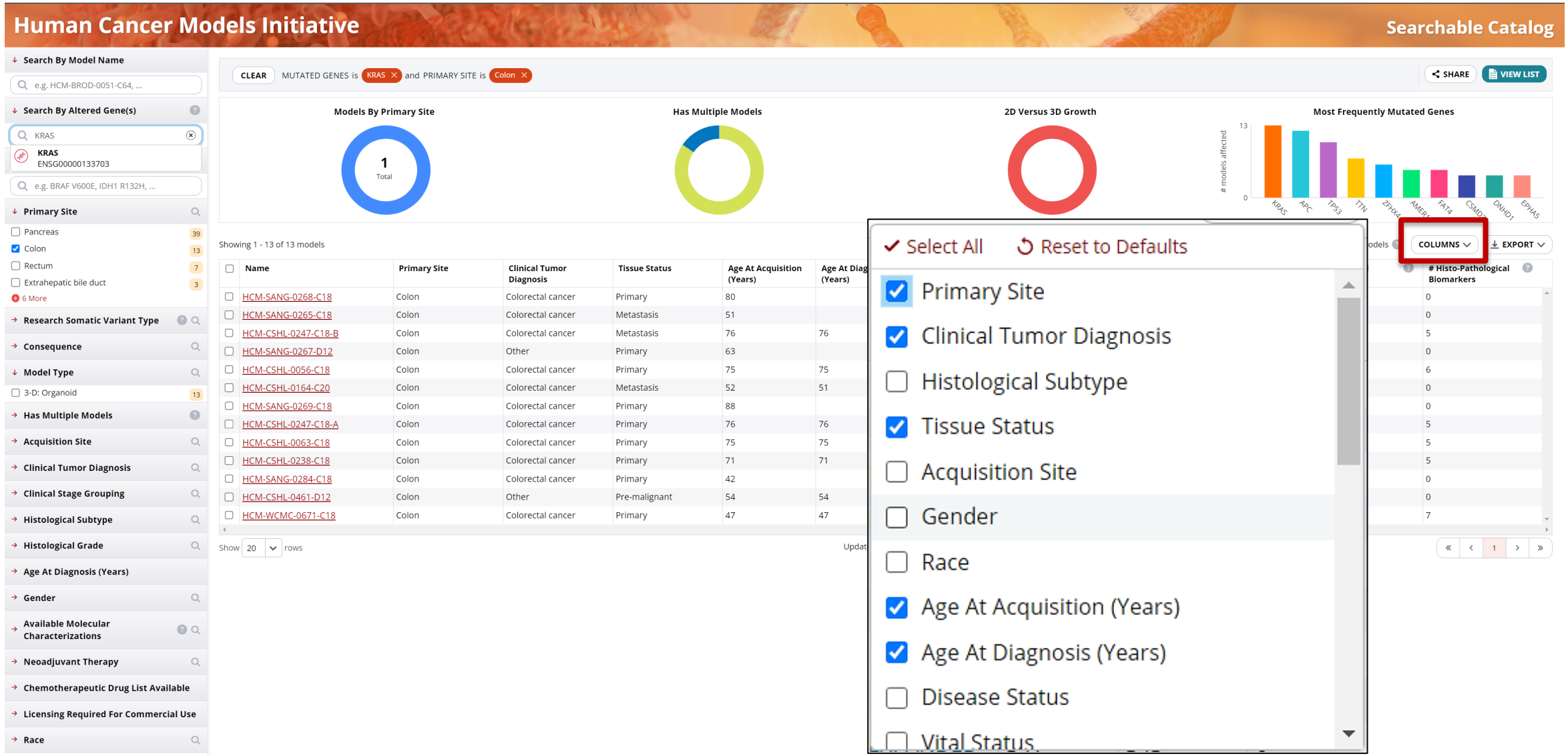
Landing Page: Catalog Features Highlights



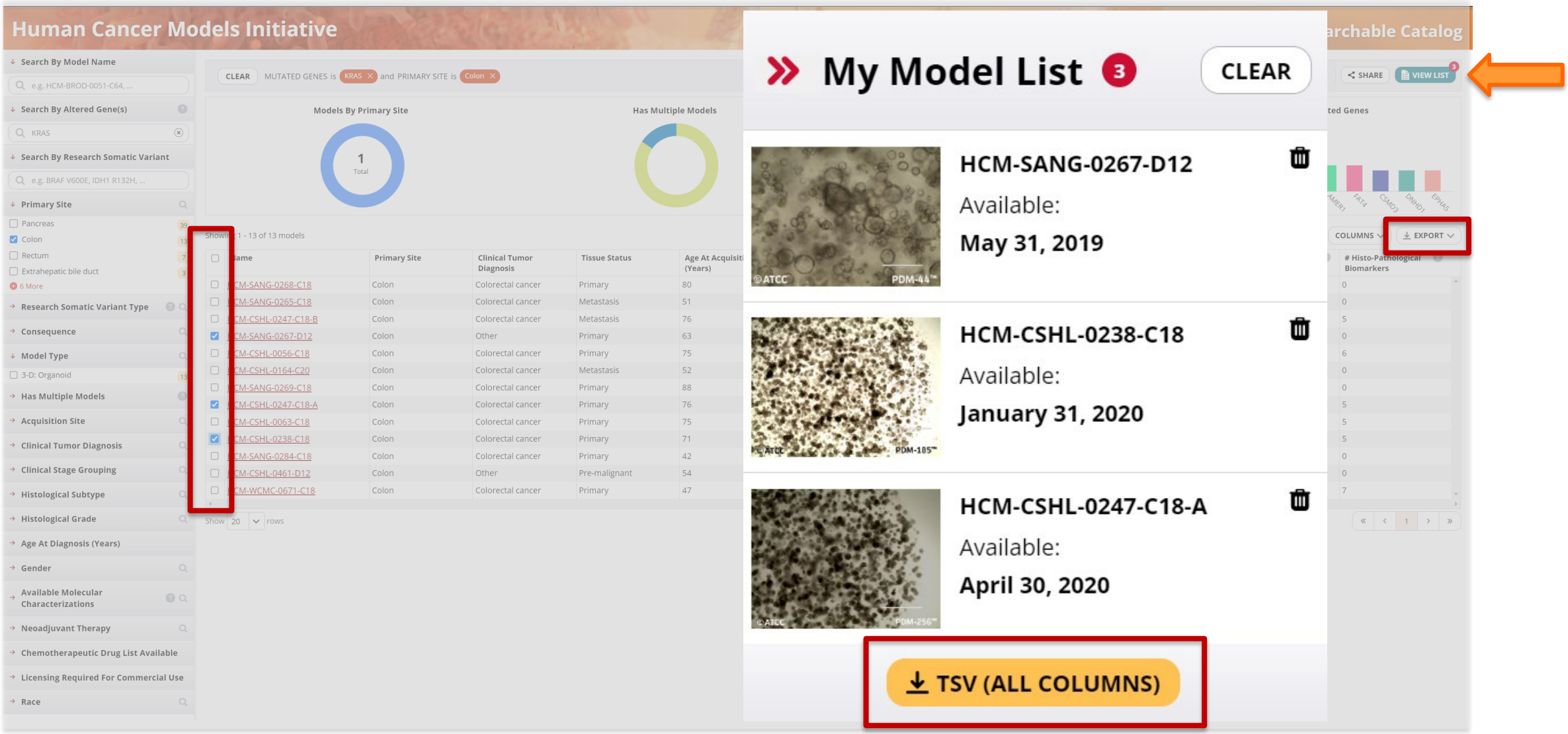
Landing Page: Filter Panel and Interactive Search Results



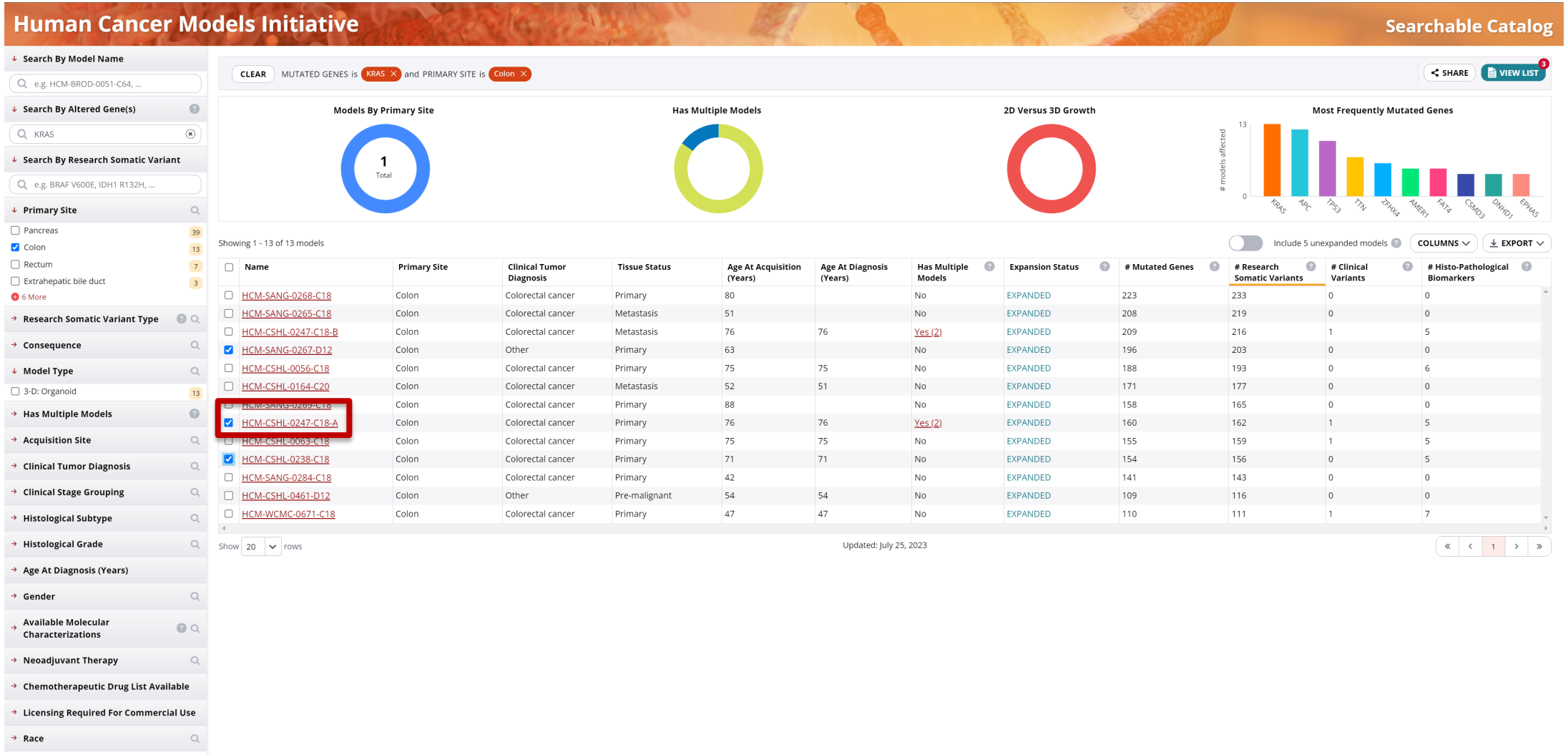
Landing Page: Customized Data Visualization and Export



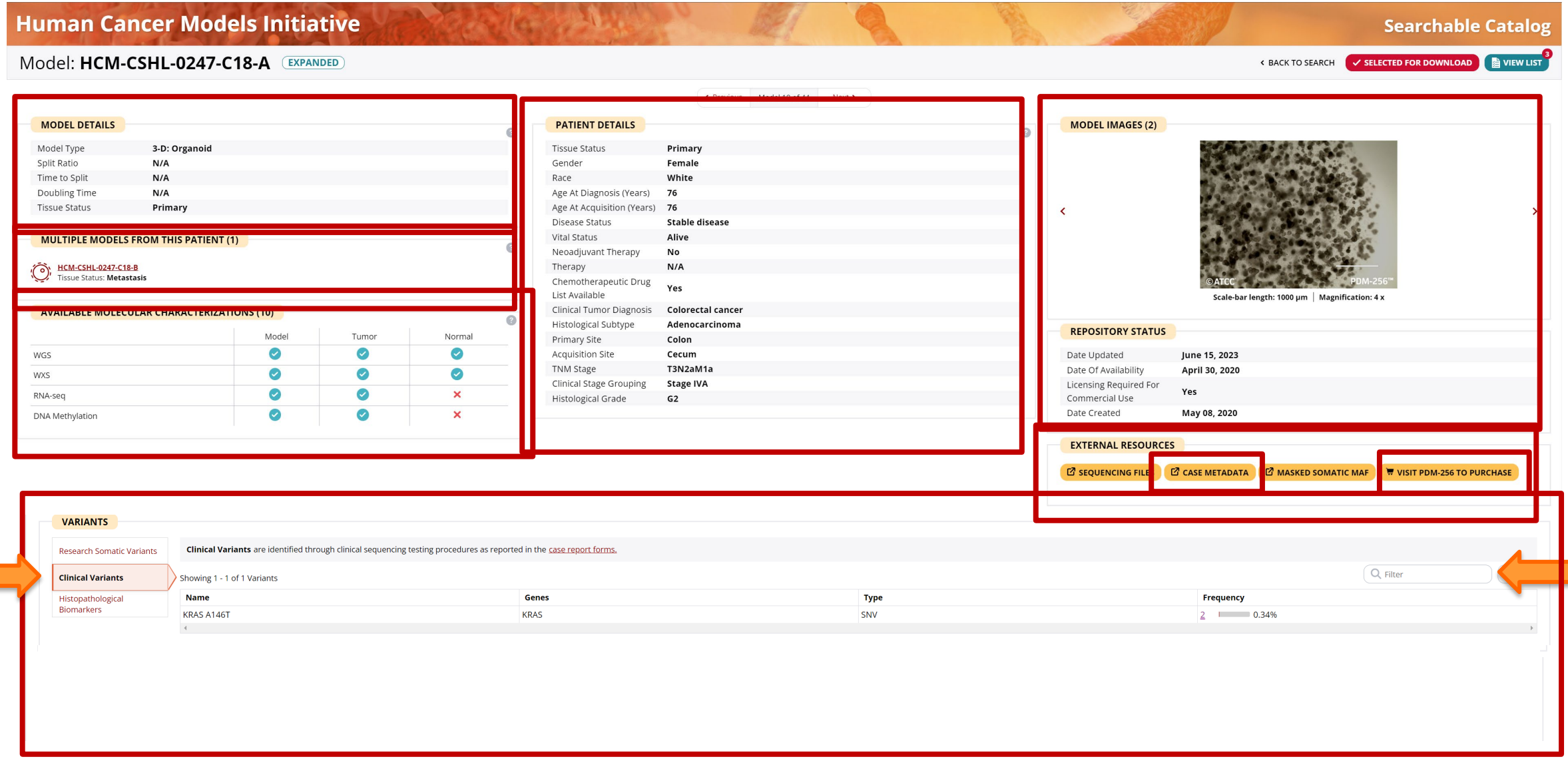
Landing Page: Customized Data Visualization and Export



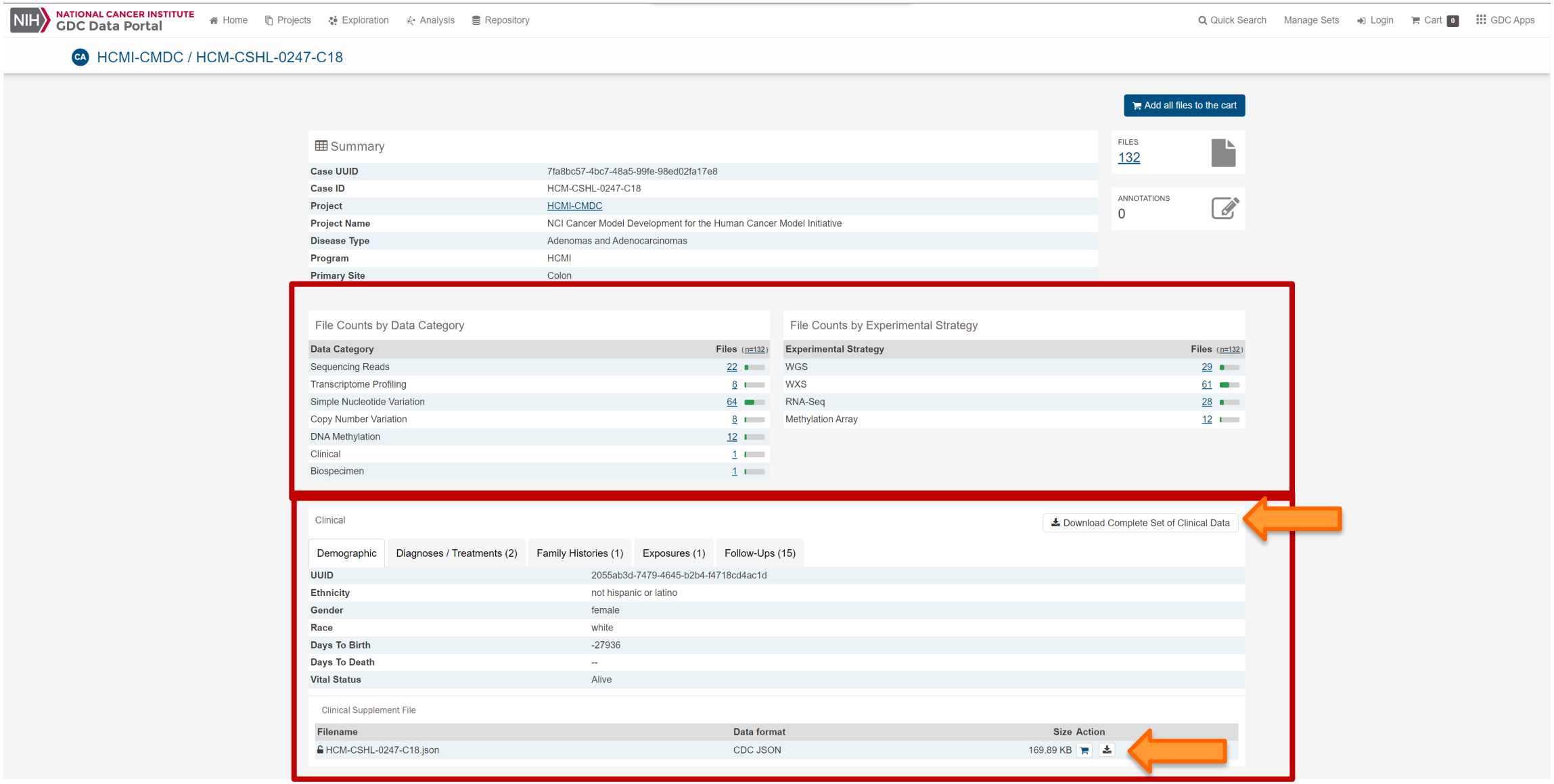
Landing Page: Navigating to Individual Model Pages



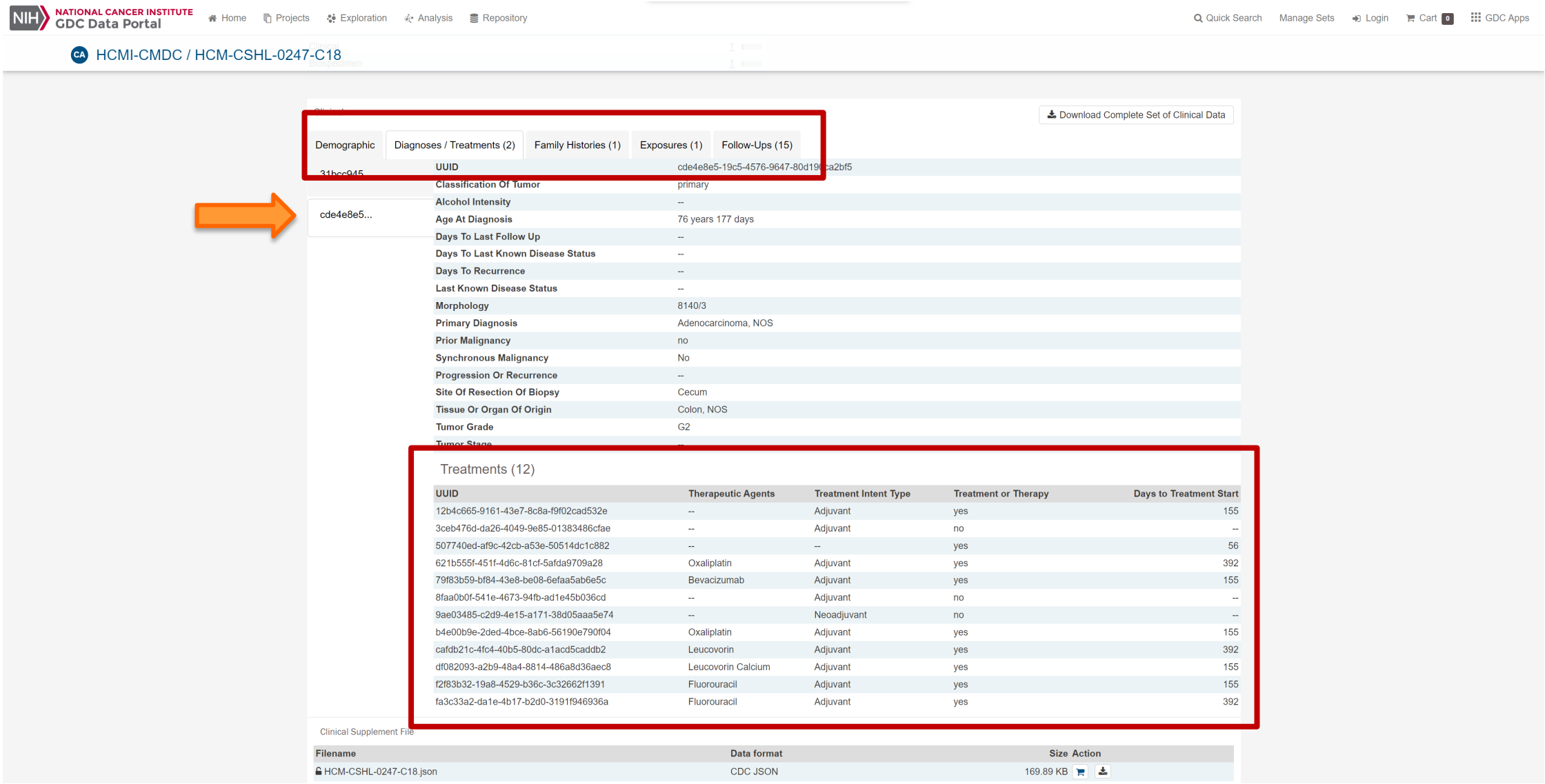
Individual Model Pages: Available Catalog Data for Selected Model



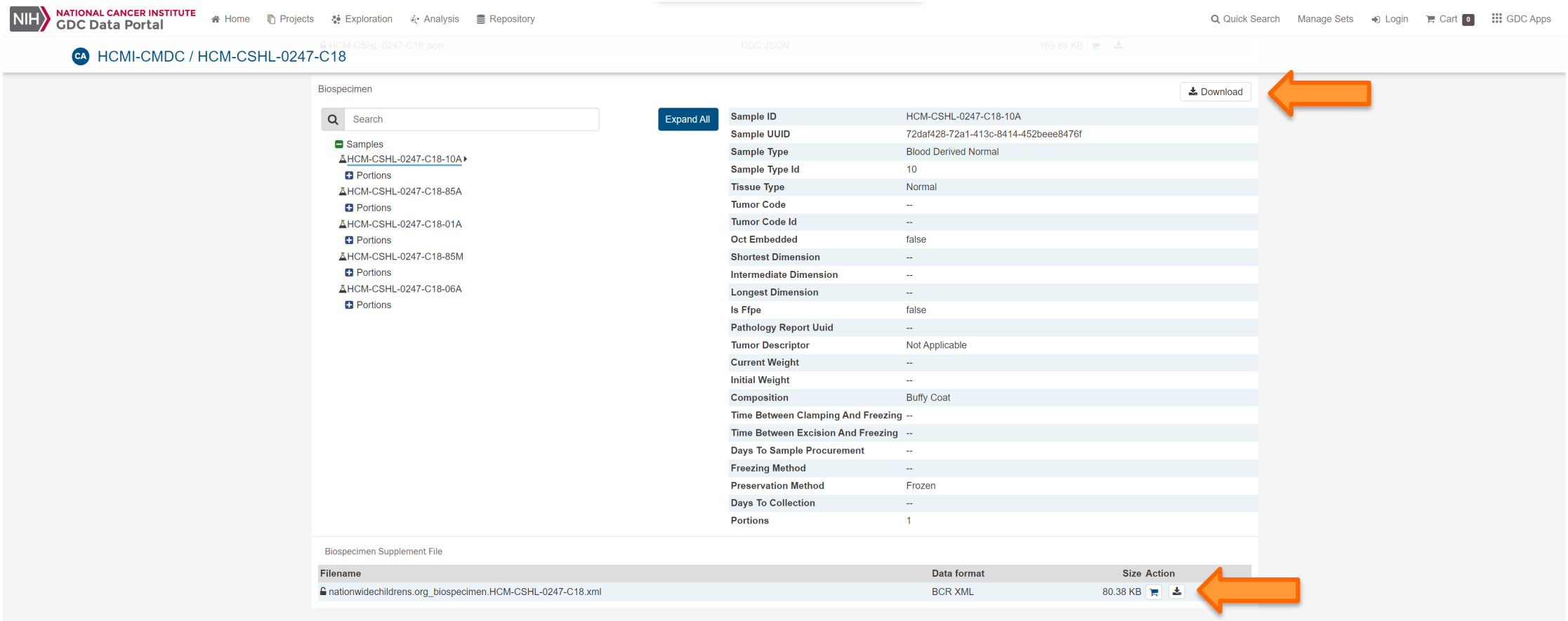
External Resources: Available Data at GDC



External Resources: Available Data at GDC



External Resources: Available Data at GDC



Acknowledgments

- **Center for Cancer Genomics**
- **Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc.**
- **NCI Contracting Office**
- **Center to Reduce Cancer Health Disparities**

HCMI Consortium Founders:

- National Cancer Institute
- Wellcome Sanger Institute
- Cancer Research UK
- Hubrecht Organoid Technology

HCMI Searchable Catalog:

- Ontario Institute for Cancer Research
- Center for Biomedical Informatics and Information Technology

Cancer Model Development Centers:

- Broad Institute
- Cold Spring Harbor Laboratory
  - University of Verona and Hubrecht Institute
- Stanford University
- Weill Cornell Medical College

CMDC Model Development Pipeline:

- Biospecimen Processing Center: Nationwide Children's Hospital
- Clinical Data Center: Information Management Systems
- NCI Data Coordinating Center
- Genome Characterization Centers: Broad Institute and University of North Carolina
- Genomic Data Commons (GDC)

HCMI Model Distributor:

- ATCC

www.cancer.gov

www.cancer.gov/espanol

HCMI Data in the Genomic Data Commons

17 August 2023

Bill Wysocki, Ph.D. – GDC User Services Lead
Center for Translational Data Science
University of Chicago

HCMI Data in the GDC

- *Brief Introduction to the GDC*
- *Available HCMI Data*
- *HCMI Data Download Walkthrough*
- *Controlled Data Access*

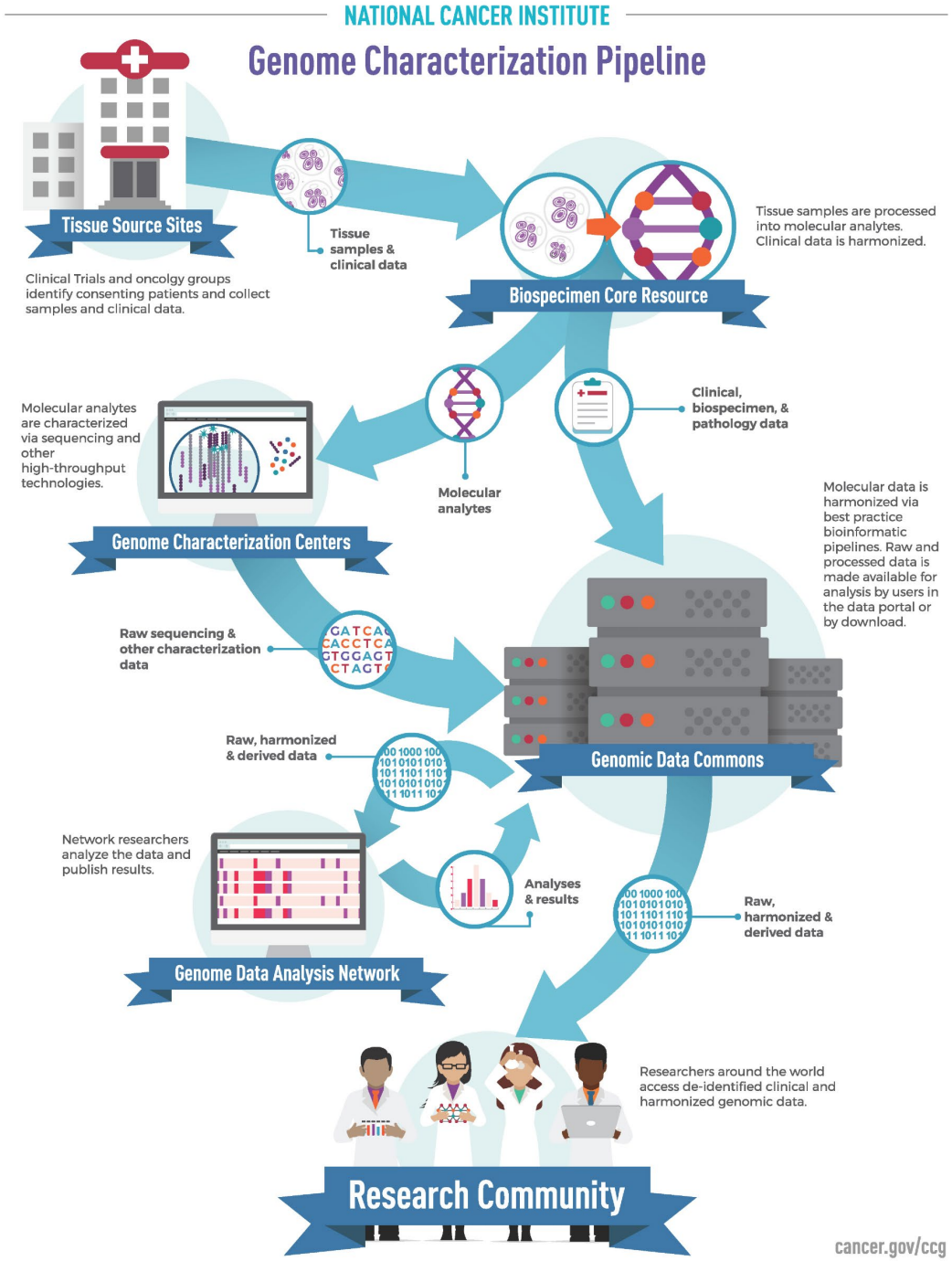
Introduction to the GDC

The Genomic Data Commons

The GDC is a data repository for cancer genomics.
This includes molecular, biospecimen, and clinical data.

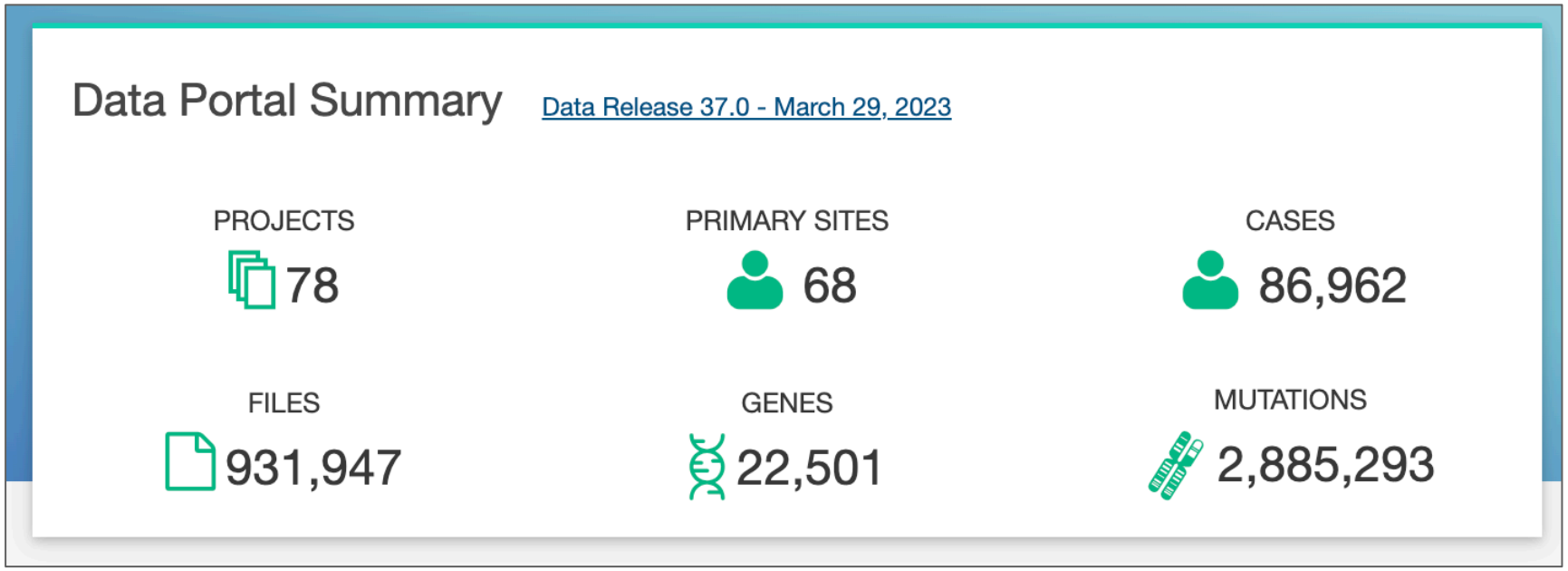
Data is processed using a set of standardized pipelines and stored in a flexible data model.

Data is submitted by researchers to be processed and eventually released to the wider research community



GDC Data Summary – August 2023 – Data Release 37

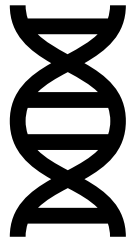
~3.8 PB of data across **932K data files**
86,962 cases across **78 projects**
2.9 million explorable somatic mutations



GDC Data

Molecular data – Includes data from multiple experimental strategies. Harmonized data starts at alignment and includes downstream derived data.

WXS, WGS, RNA-Seq, Methylation

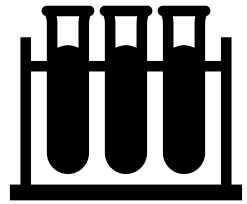


Clinical data – Uploaded as metadata associated with cases (patients) but is also available as supplement files.

Demographic, Diagnosis, Exposure, Family History, Pathology

Biospecimen data – Also uploaded as metadata, but is associated with samples, portions, analytes, and aliquots. Also available as supplement files.

Sample, Portion, Analyte, Aliquot, Slides

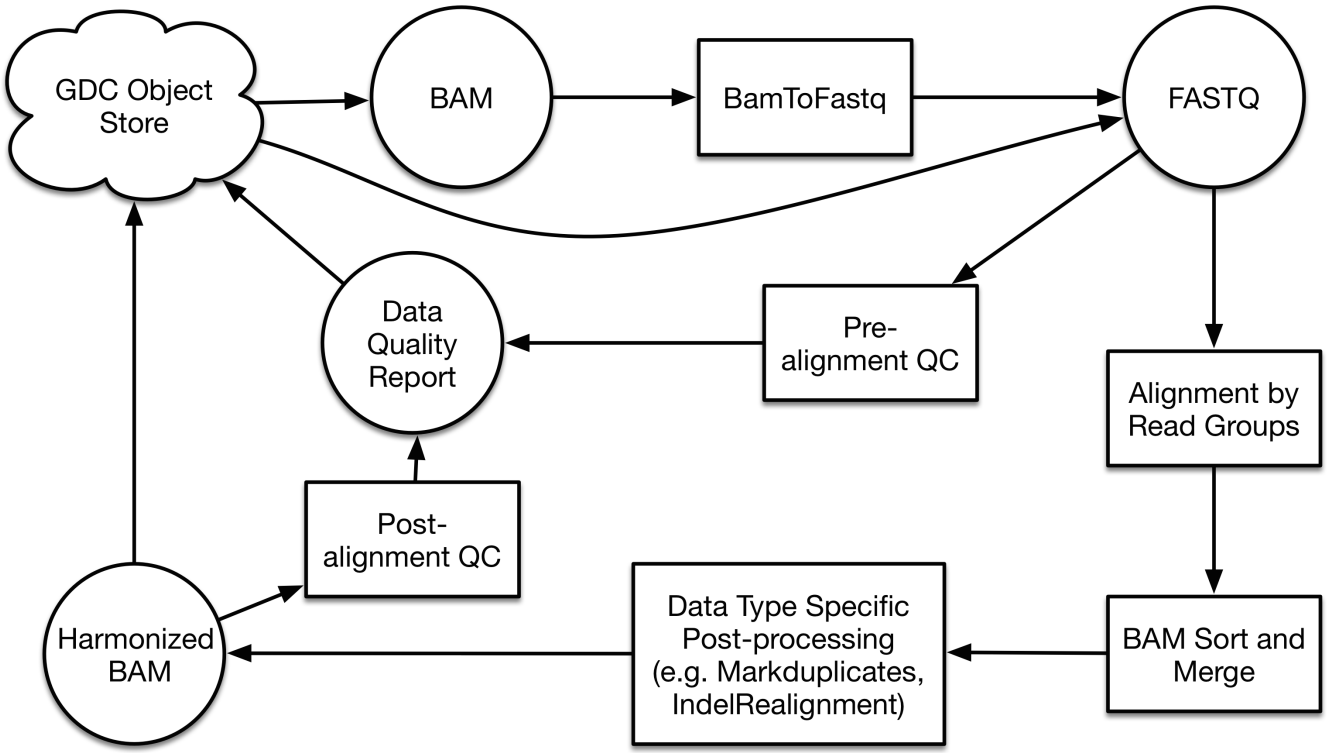


HCMI Data Available in the GDC

Harmonized Data

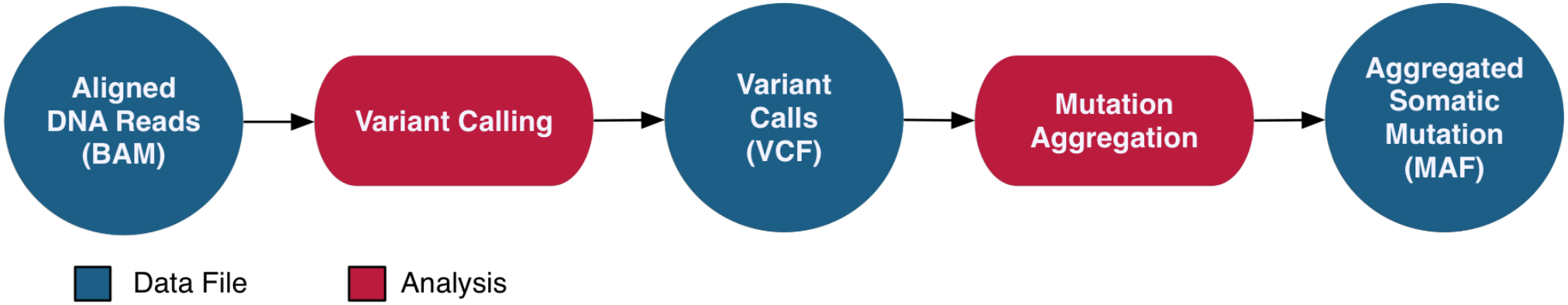
Molecular Data Strategies Available for HCMI

- Alignments available for all NGS strategies
  - All samples aligned to the same reference genome



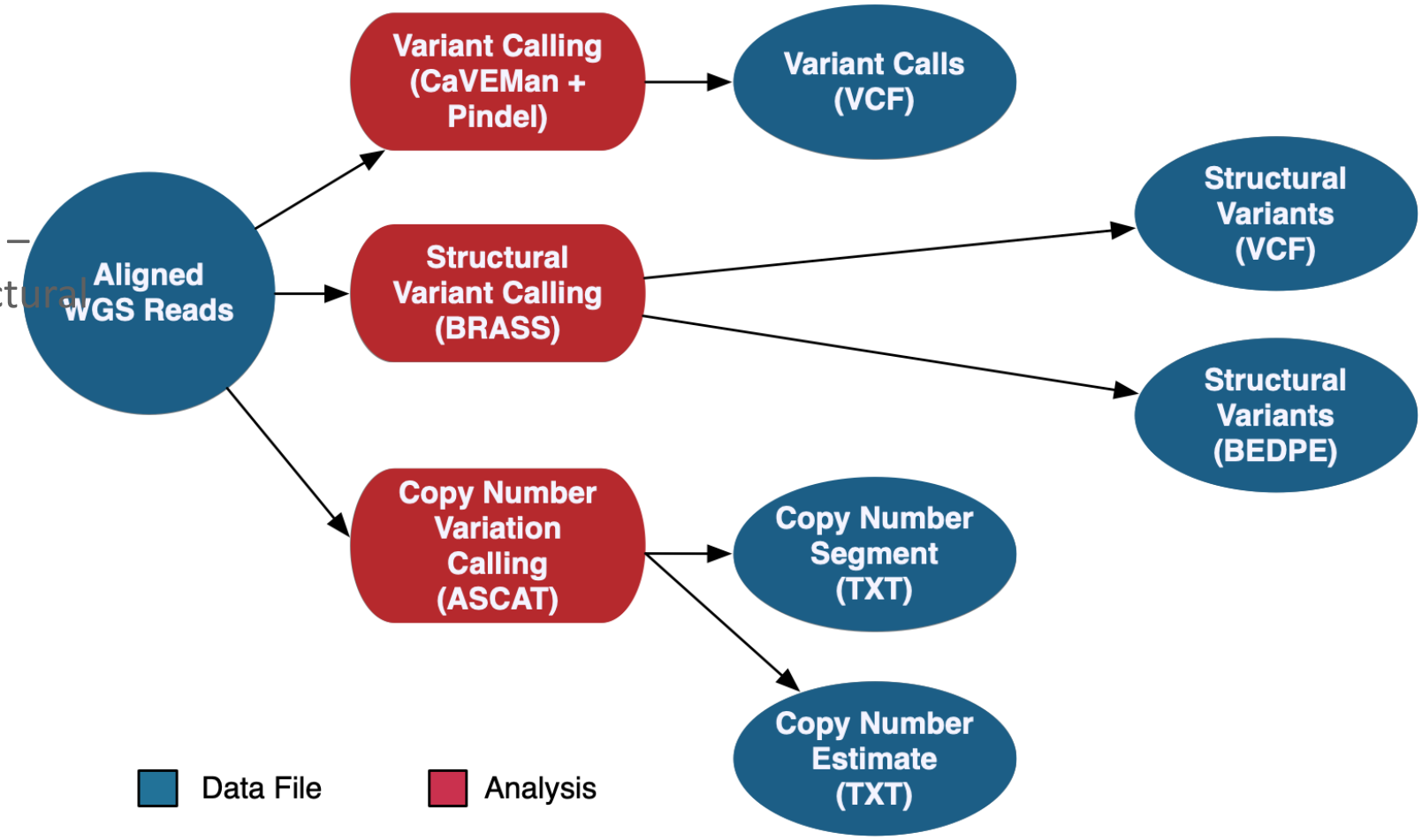
Molecular Data Strategies Available for HCMI

- **Whole Exome Sequencing**
  - Somatic variant calling pipeline – SNVs and indels
  - Available raw or filtered



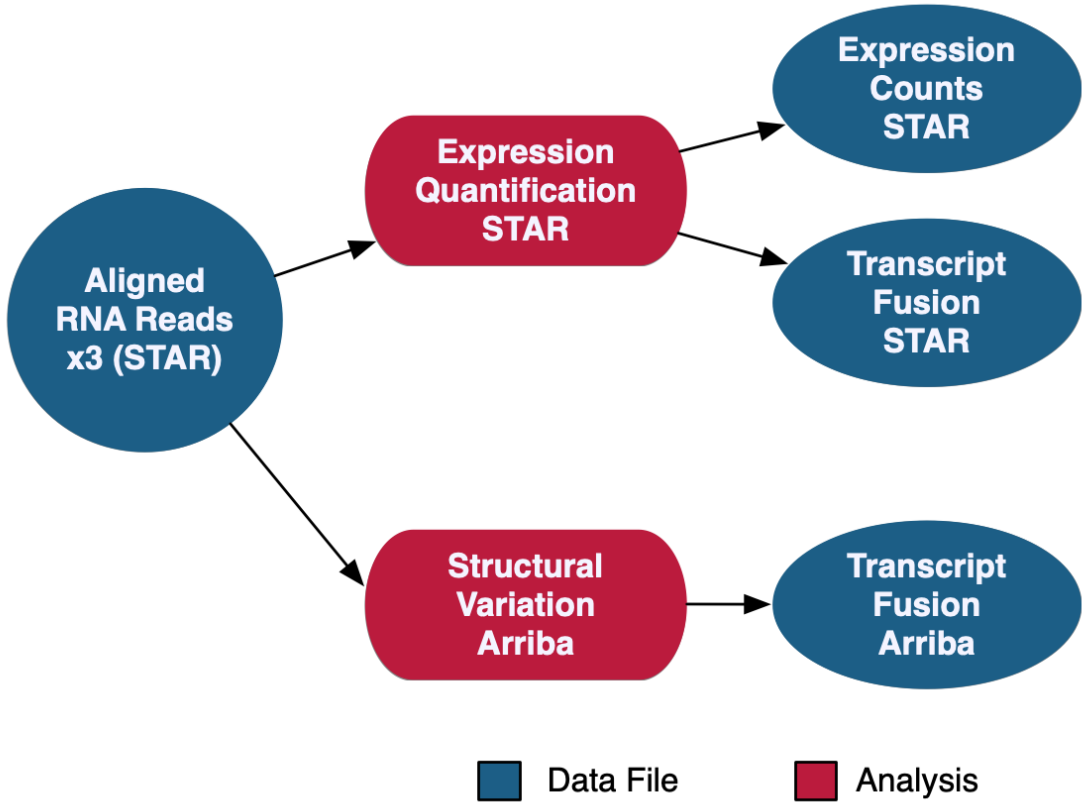
Molecular Data Strategies Available for HCMI

- **Whole Genome Sequencing**
  - Somatic variant calling pipelines – SNVs, indels, copy number, structural



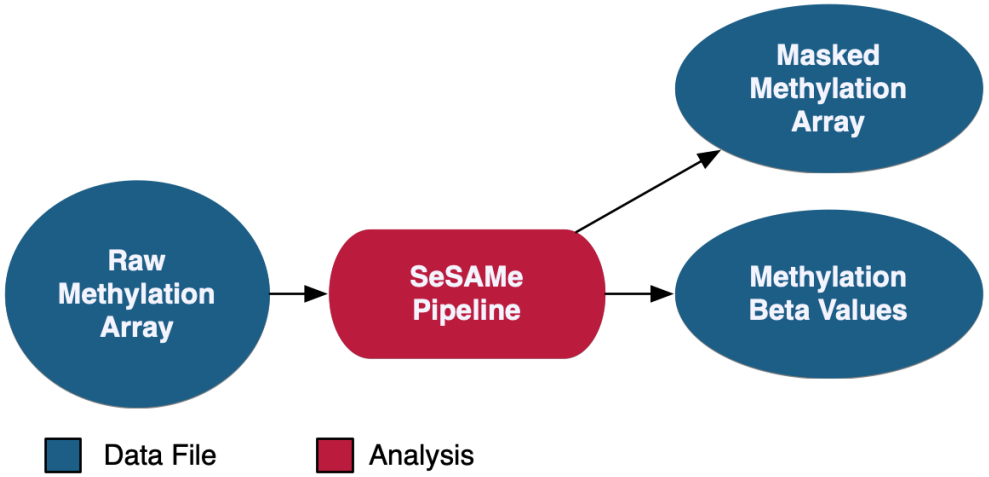
Molecular Data Strategies Available for HCMI

- **RNA-Seq**
  - Gene expression values available – Read counts per gene and normalized.
  - Transcript Fusion and Splice Junction files are also available.



Molecular Data Strategies Available for HCMI

- **Methylation Array**
  - Raw array files available with germline probes removed.
  - Methylation Beta Values are also available.



HCMI Data Download Walkthrough

Use Case: *Downloading RNA-Seq Gene Expression Data*

HCMI Case Page

HCMI-CMDC / HCM-CSHL-0247-C18

Add all files to the cart

FILES
132

ANNOTATIONS
0

Summary

Case UUID	7fa8bc57-4bc7-48a5-99fe-98ed02fa17e8
Case ID	HCM-CSHL-0247-C18
Project	HCMI-CMDC
Project Name	NCI Cancer Model Development for the Human Cancer Model Initiative
Disease Type	Adenomas and Adenocarcinomas
Program	HCMI
Primary Site	Colon

File Counts by Data Category

Data Category	Files (n=132)
Sequencing Reads	22
Transcriptome Profiling	8
Simple Nucleotide Variation	64
Copy Number Variation	8
DNA Methylation	12
Clinical	1
Biospecimen	1

File Counts by Experimental Strategy

Experimental Strategy	Files (n=132)
WGS	29
WXS	61
RNA-Seq	28
Methylation Array	12

HCMI Case Page

HCMI-CMDC / HCM-CSHL-0247-C18

Add all files to the cart

Summary

Case UUID	7fa8bc57-4bc7-48a5-99fe-98ed02fa17e8
Case ID	HCM-CSHL-0247-C18
Project	HCMI-CMDC
Project Name	NCI Cancer Model Development for the Human Cancer Model Initiative
Disease Type	Adenomas and Adenocarcinomas
Program	HCMI
Primary Site	Colon

FILES
132

ANNOTATIONS
0

RNA-Seq 28

File Counts by Data Category

Data Category	Files (n=132)
Sequencing Reads	22
Transcriptome Profiling	8
Simple Nucleotide Variation	64
Copy Number Variation	8
DNA Methylation	12
Clinical	1
Biospecimen	1

File Counts by Experimental Strategy

Experimental Strategy	Files (n=132)
WGS	29
WXS	61
RNA-Seq	28
Methylation Array	12

HCMI Case Page

HCMI-CMDC / HCM-CSHL-0247-C18

Add all files to the cart

Summary

Case UUID	7fa8bc57-4bc7-48a5-99fe-98ed02fa17e8
Case ID	HCM-CSHL-0247-C18
Project	HCMI-CMDC
Project Name	NCI Cancer Model Development for the Human Cancer Model Initiative
Disease Type	Adenomas and Adenocarcinomas
Program	HCMI
Primary Site	Colon

FILES
132

ANNOTATIONS
0

RNA-Seq 28

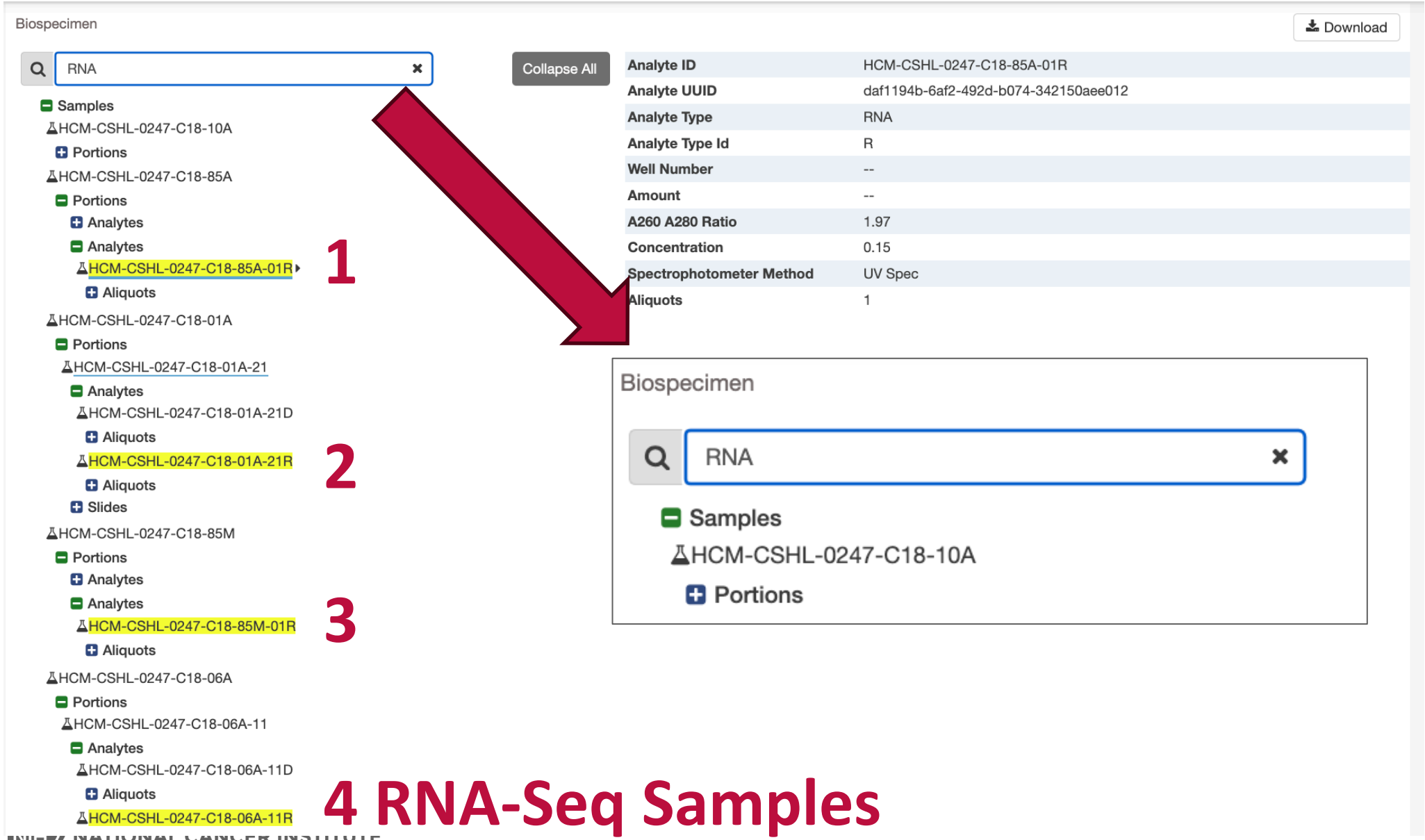
File Counts by Data Category

Data Category	Files (n=132)
Sequencing Reads	22
Transcriptome Profiling	8
Simple Nucleotide Variation	64
Copy Number Variation	8
DNA Methylation	12
Clinical	1
Biospecimen	1

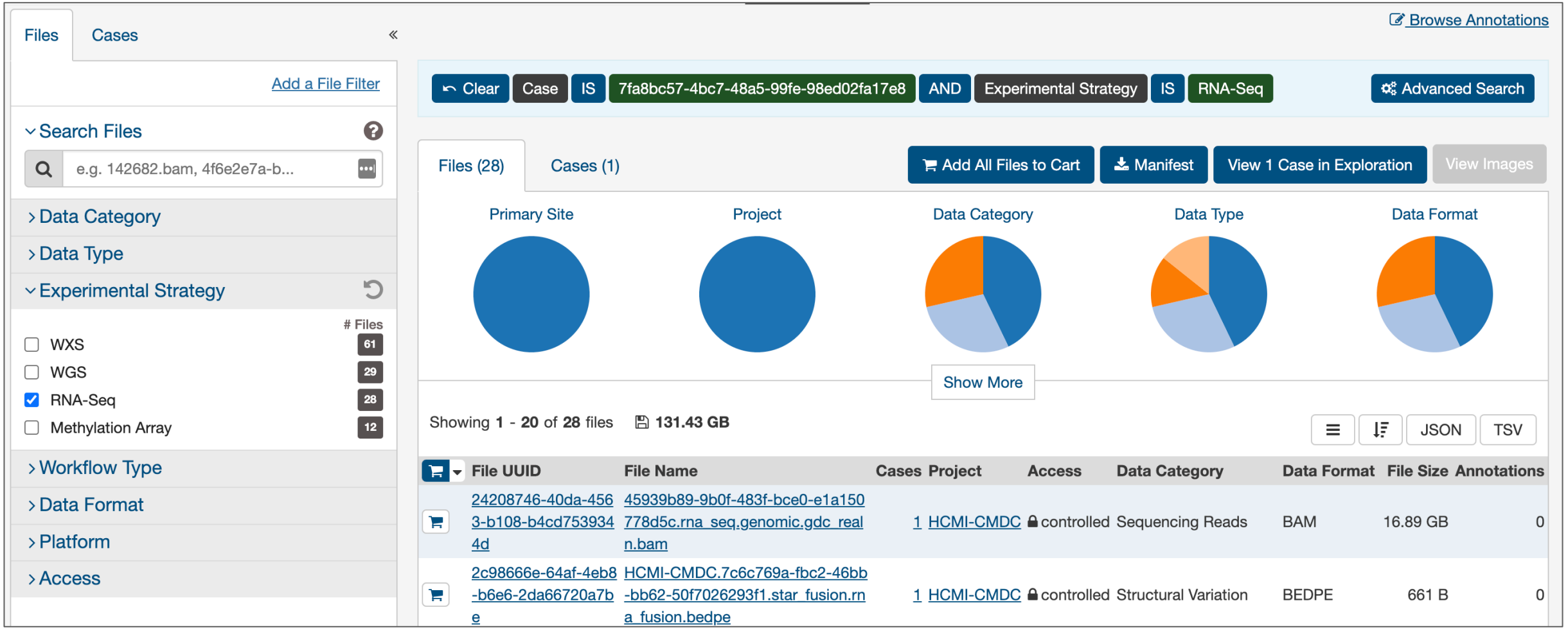
File Counts by Experimental Strategy

Experimental Strategy	Files (n=132)
WGS	29
WXS	61
RNA-Seq	28
Methylation Array	12

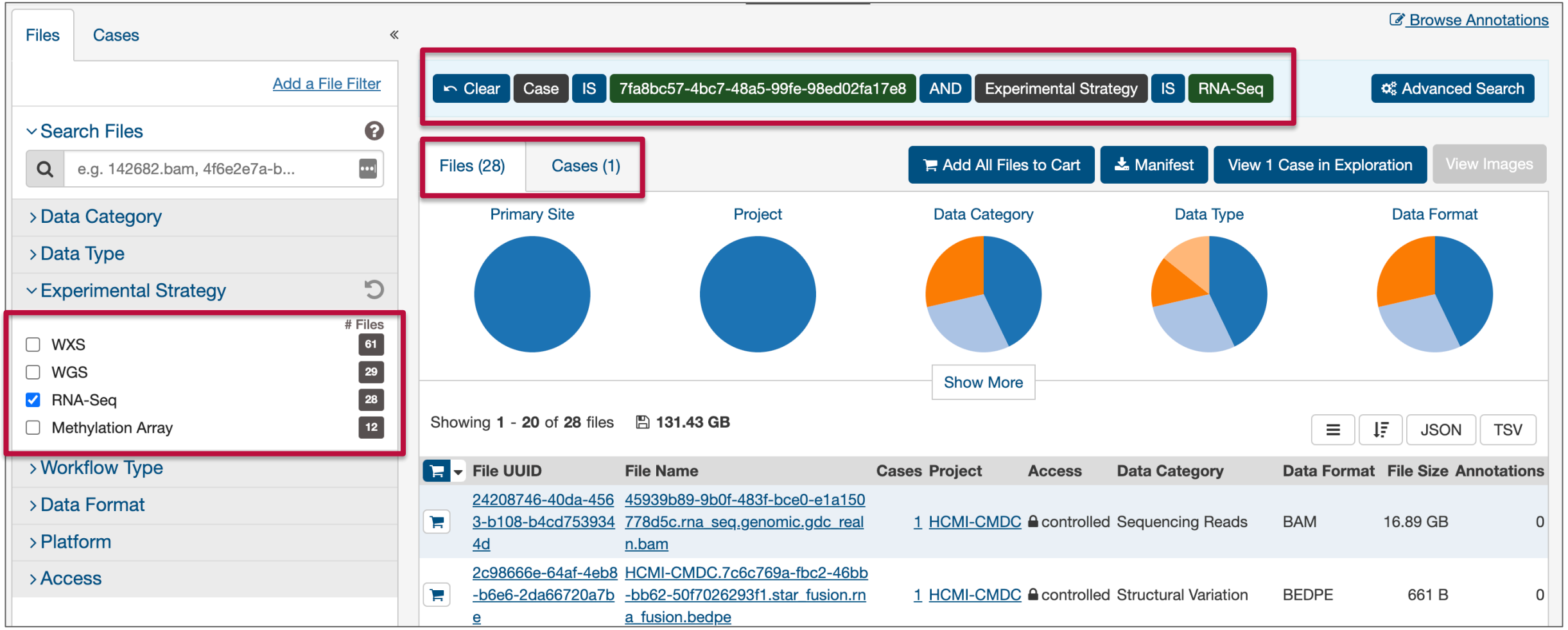
HCMI Biospecimen Check



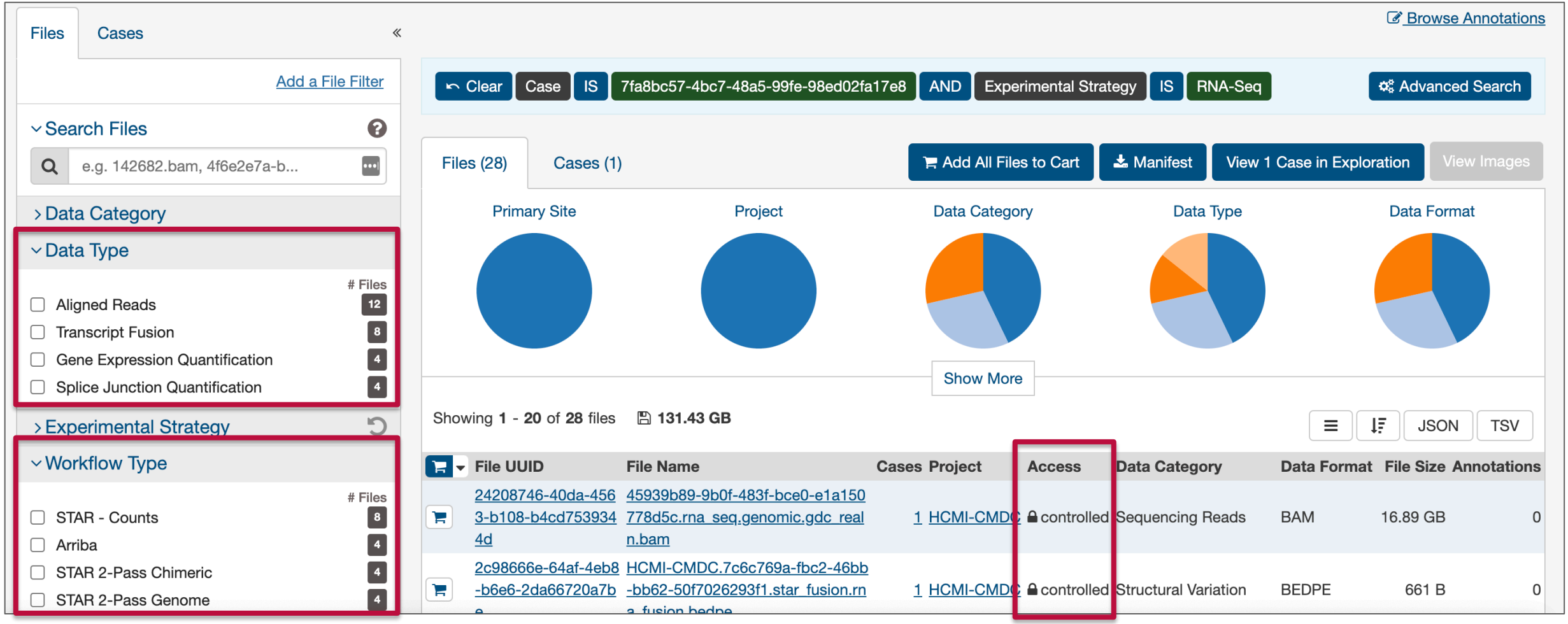
Repository View: Files associated with Case



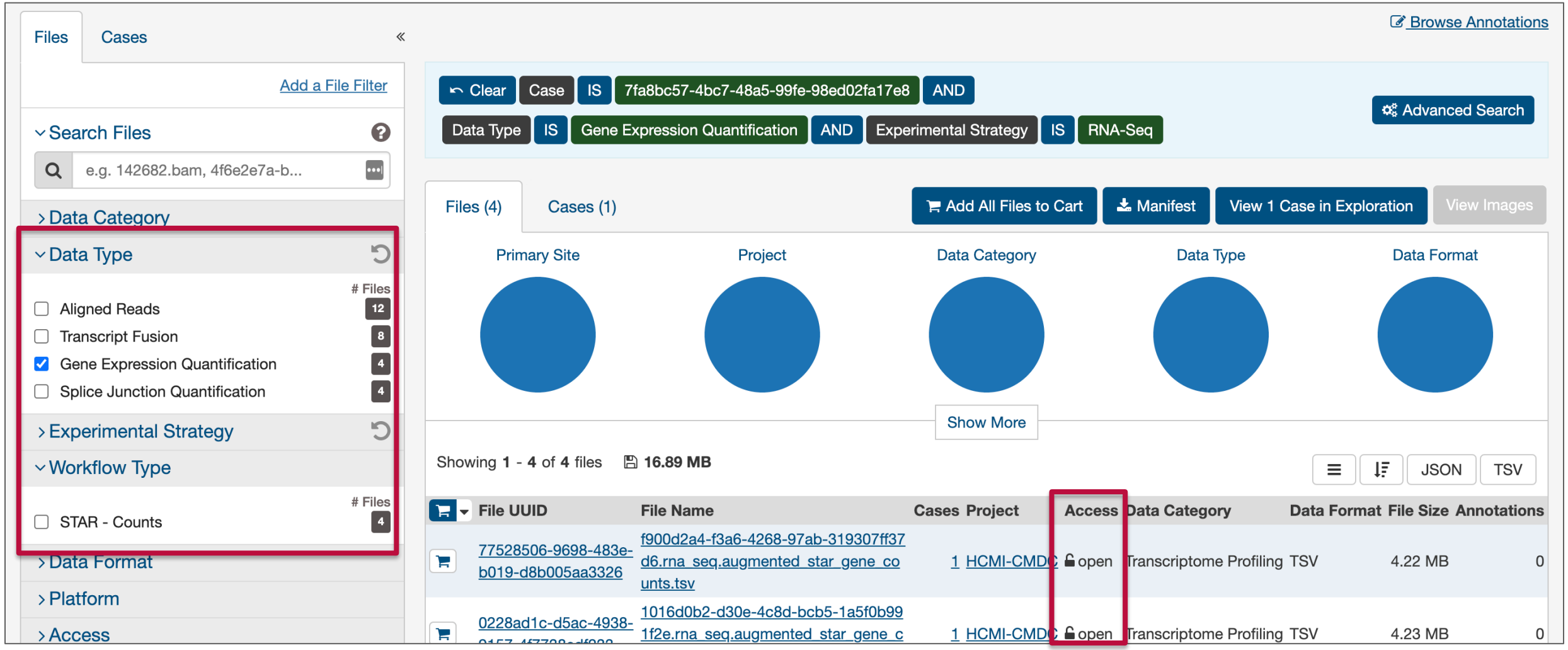
Repository View: Files associated with Case



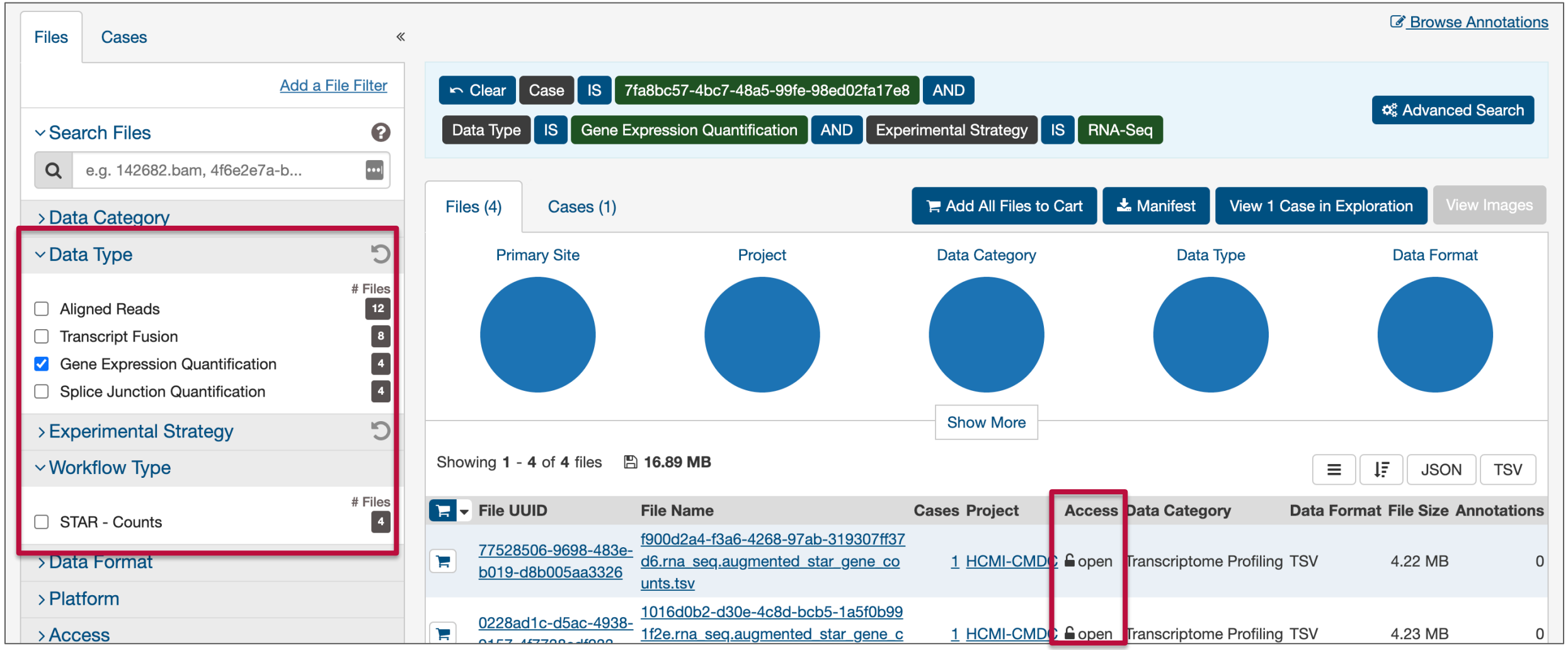
Repository View: More Filters



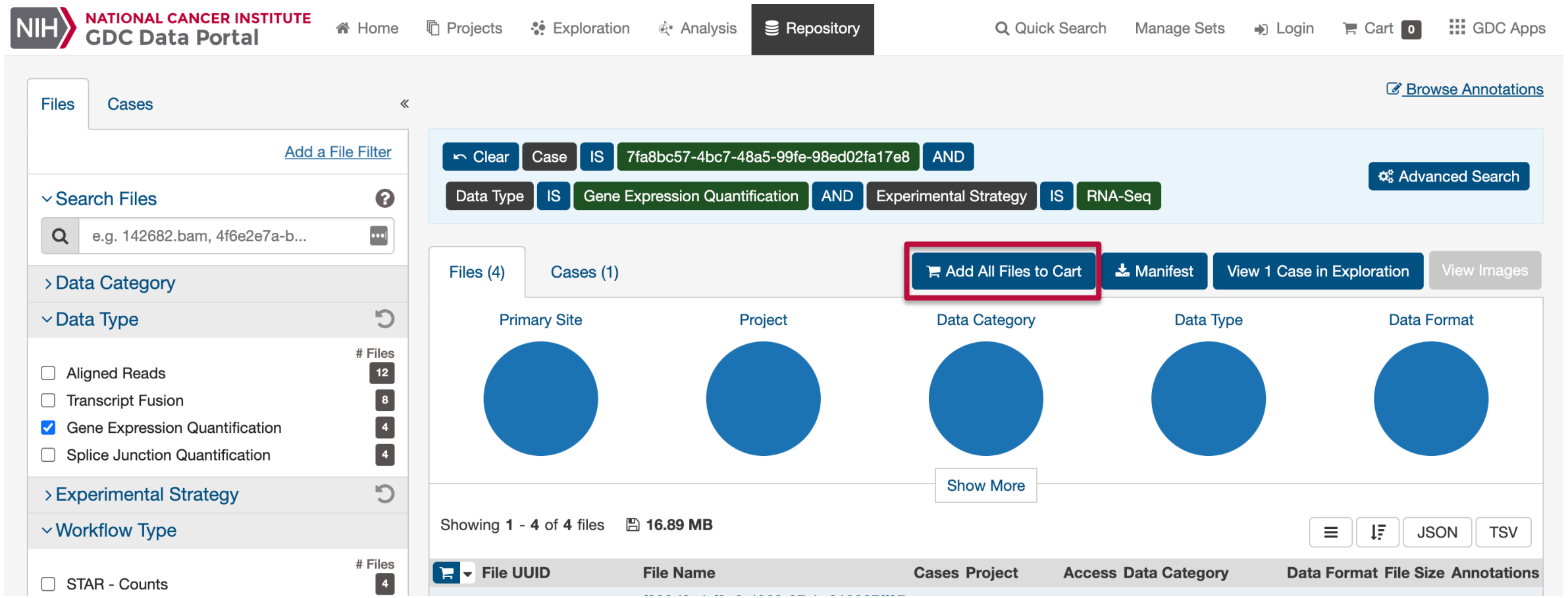
Repository View: More Filters



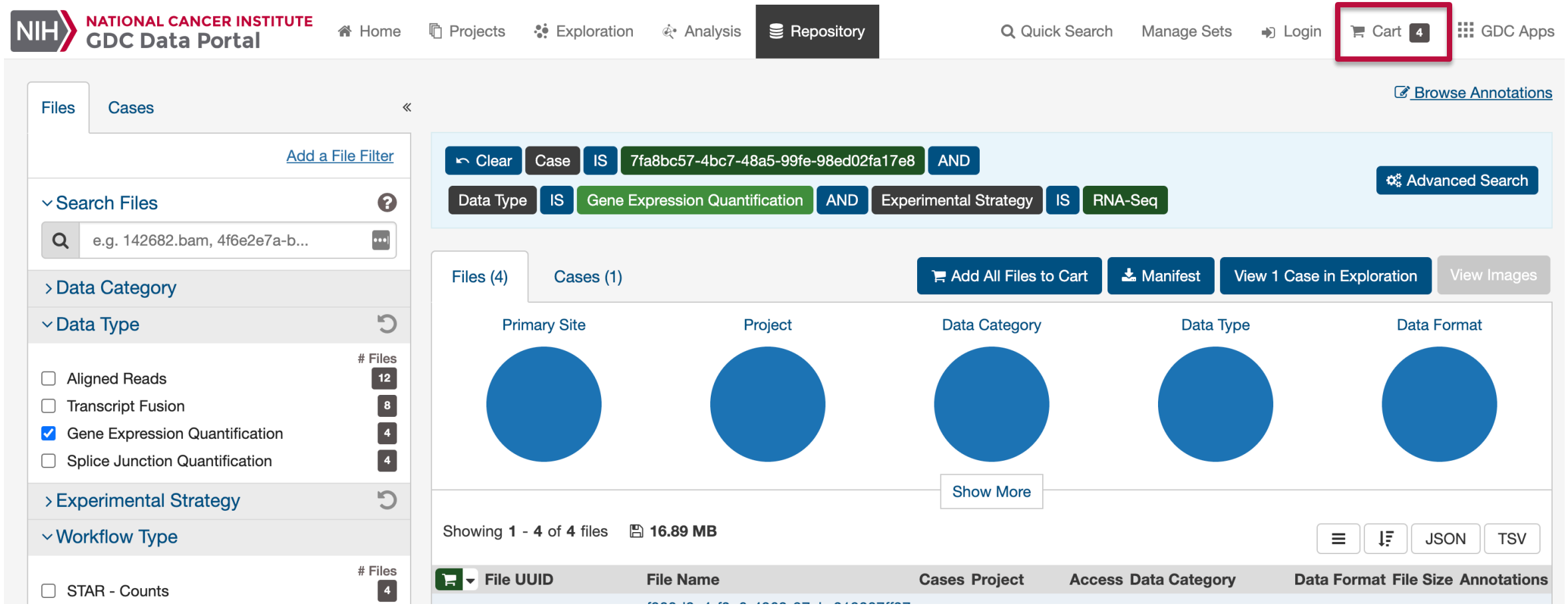
Repository View: More Filters



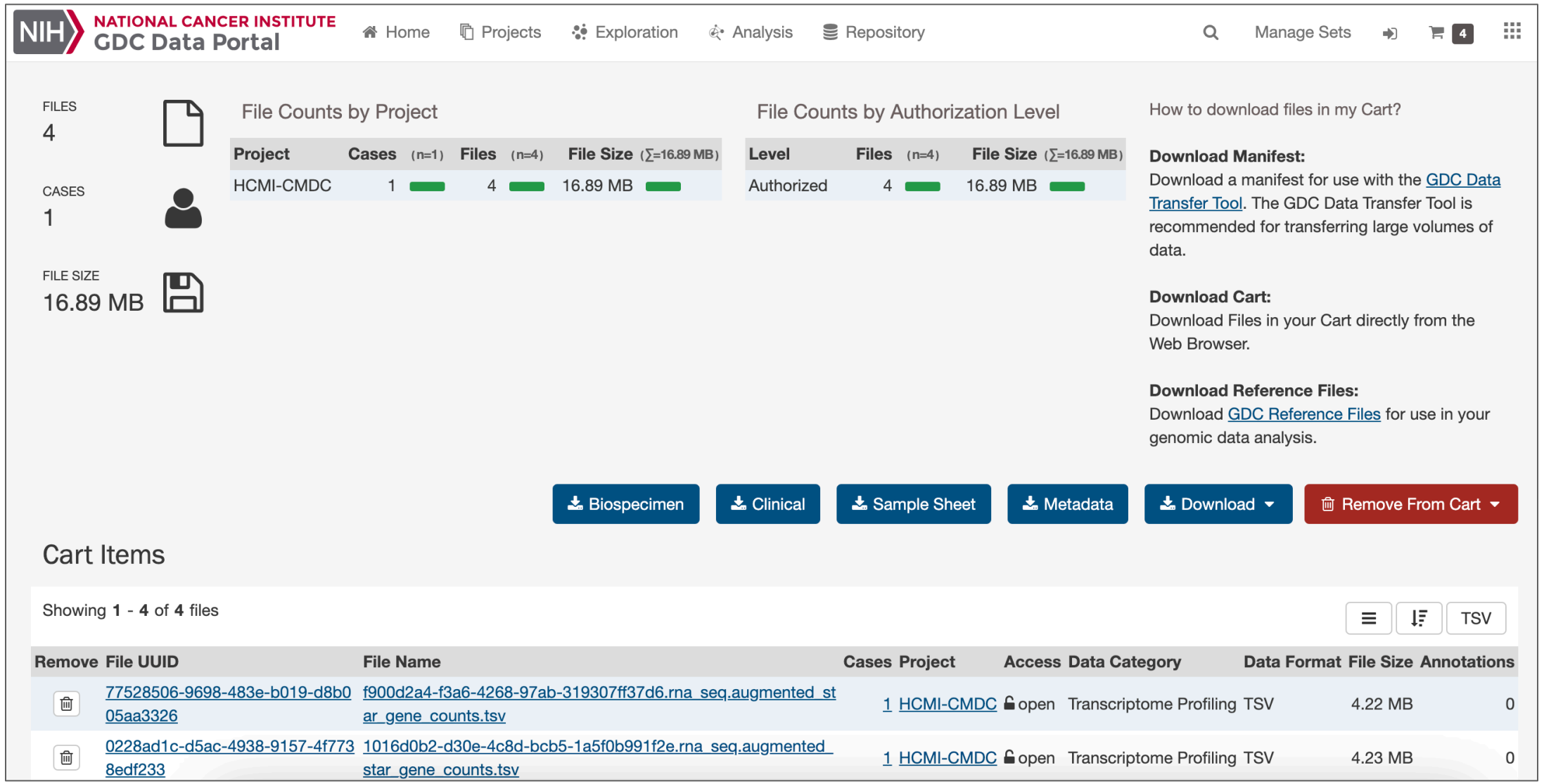
Repository View: Add Files to Cart



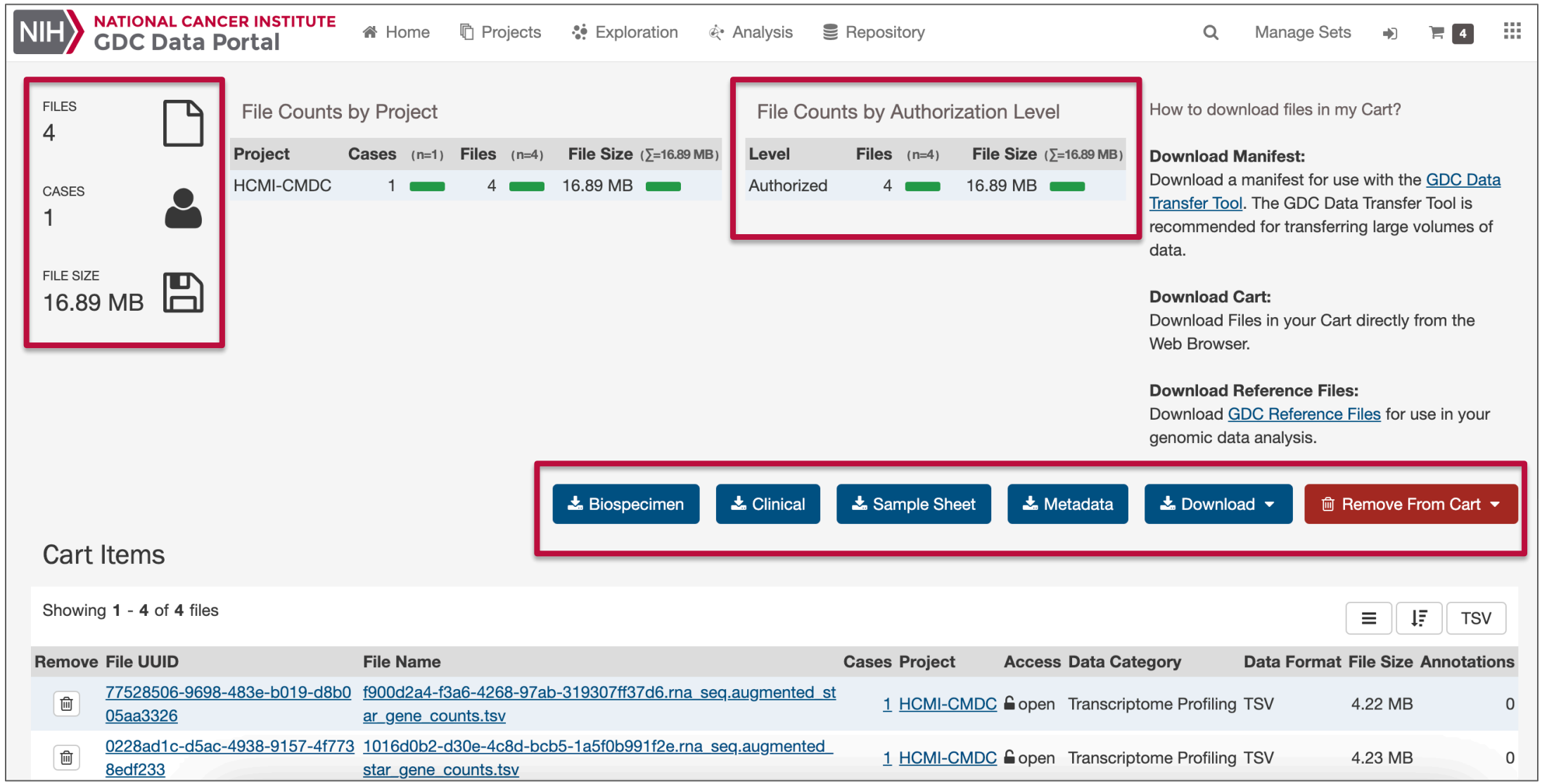
Repository View: Add Files to Cart



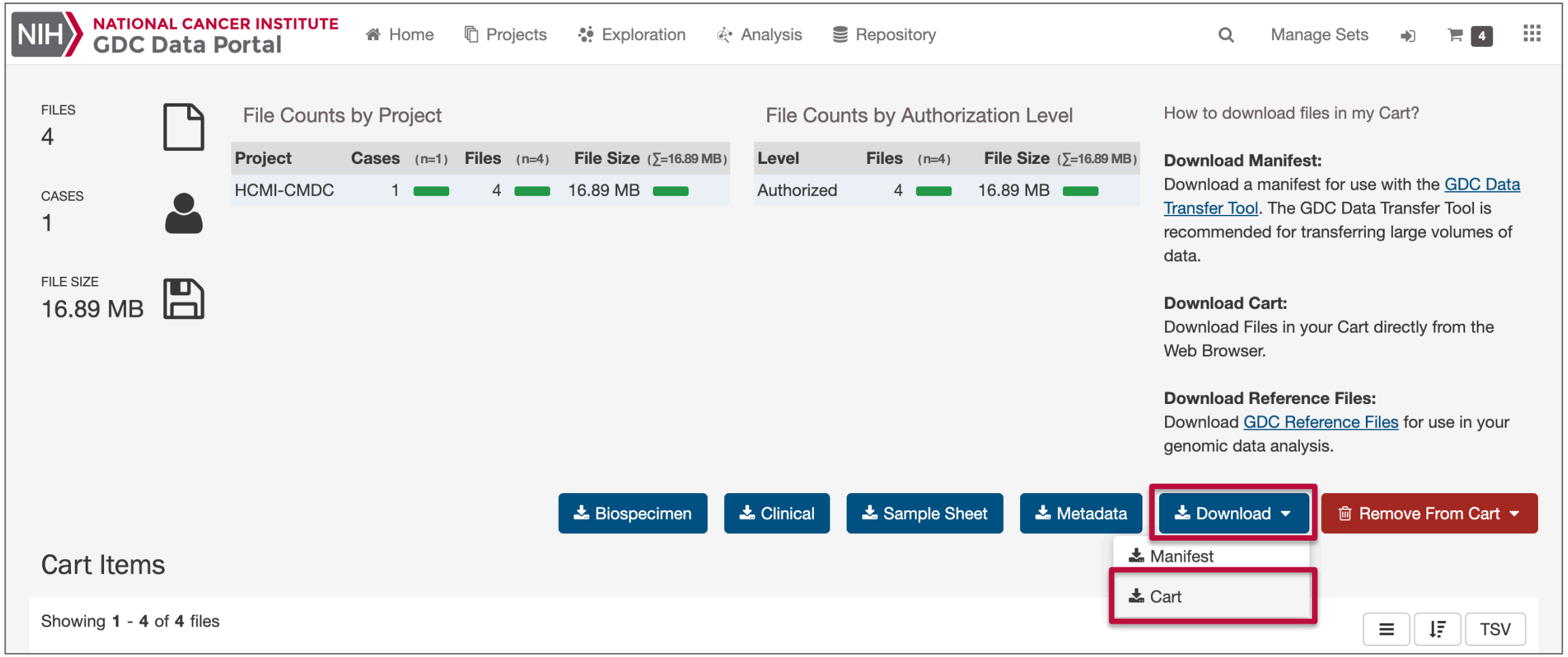
Cart: Download Files



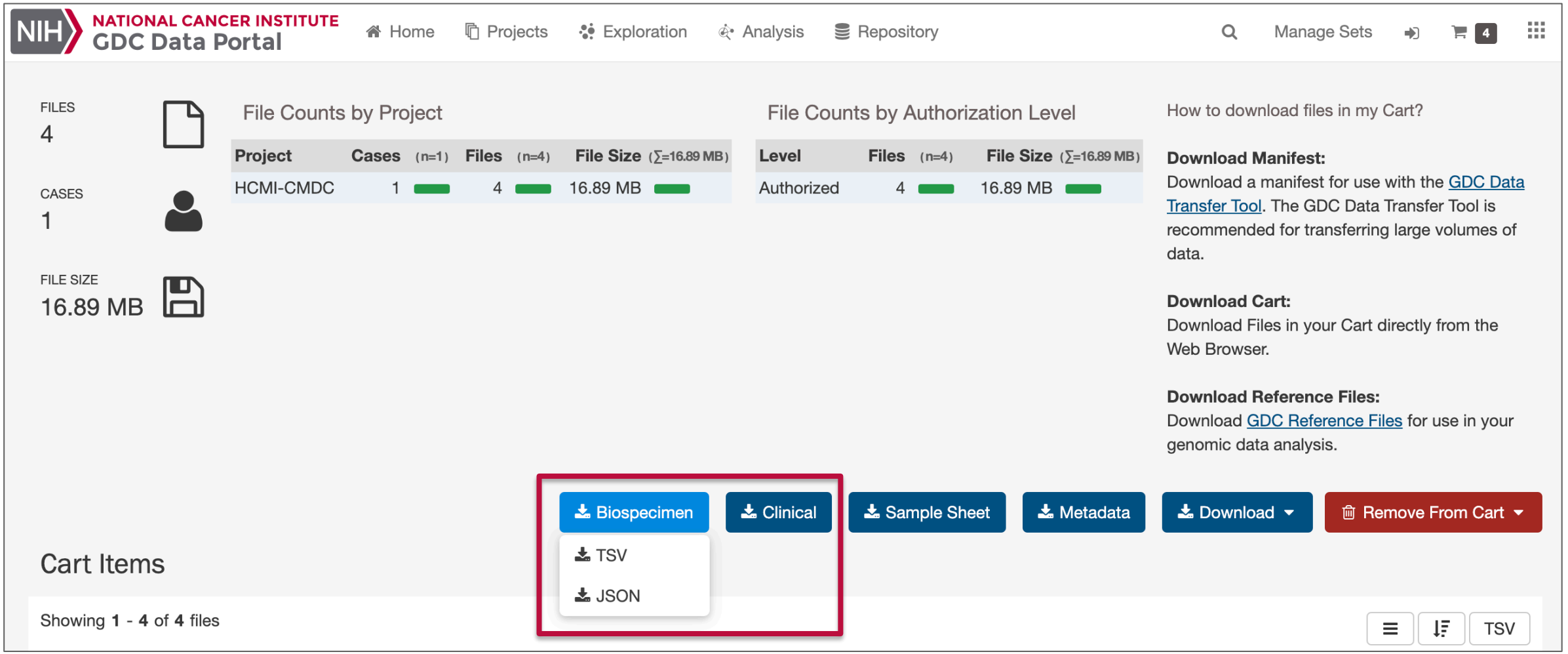
Cart: Download Files



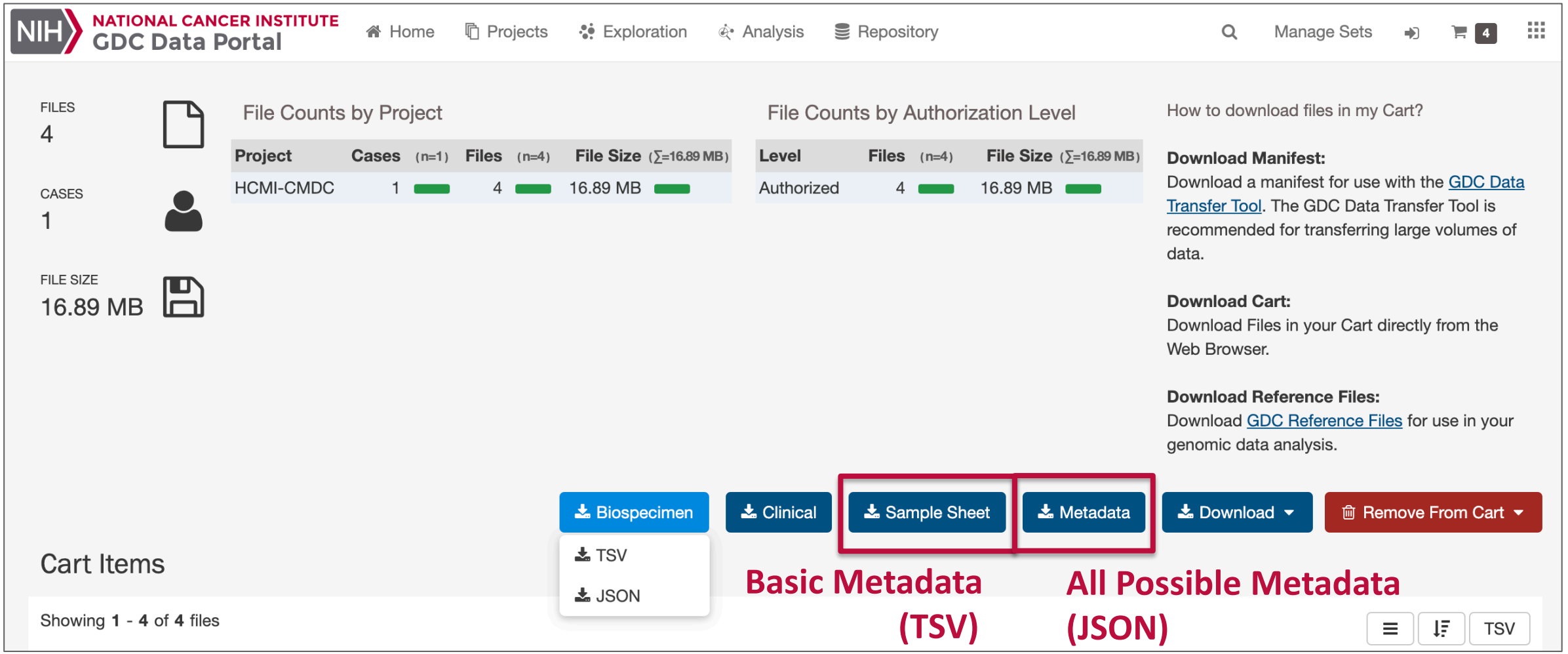
Cart: Download Files



Cart: Download File Metadata – Clinical & Biospecimen

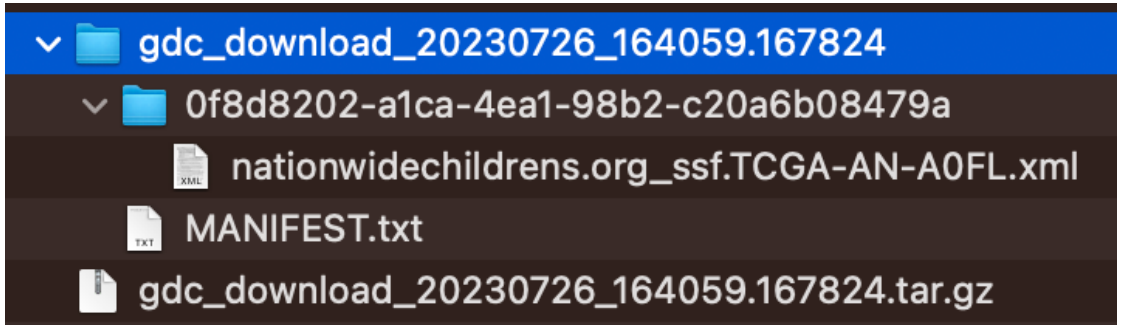


Cart: Download File Metadata – Other

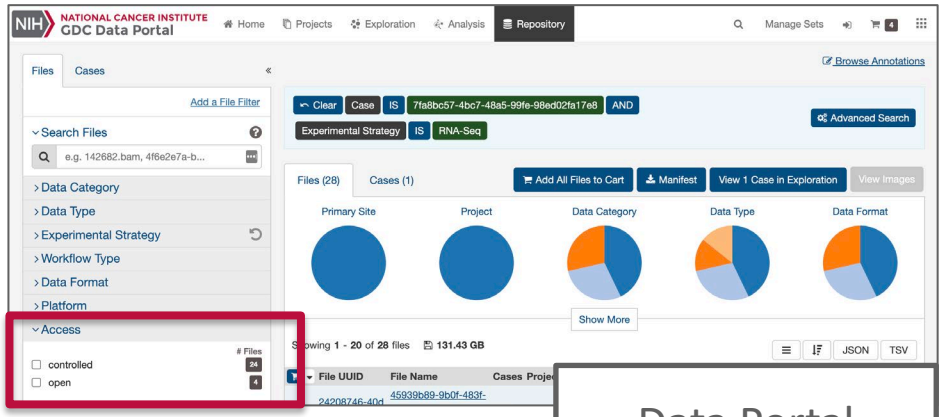


Opening Downloaded Files

File will download as a TAR.GZ file
MAC OS will open these files automatically upon download
Windows will require another program
    Two most common are **WinRAR** and **7-Zip**
When cart is downloaded, each file will be in a folder named after its UUID.
Will also contain a manifest and any annotations associate with files.



Two Additional Download Methods



Data Portal

DTT

GDC API

Reminder: Download of controlled-access files requires dbGaP access and login through the GDC Data Portal.

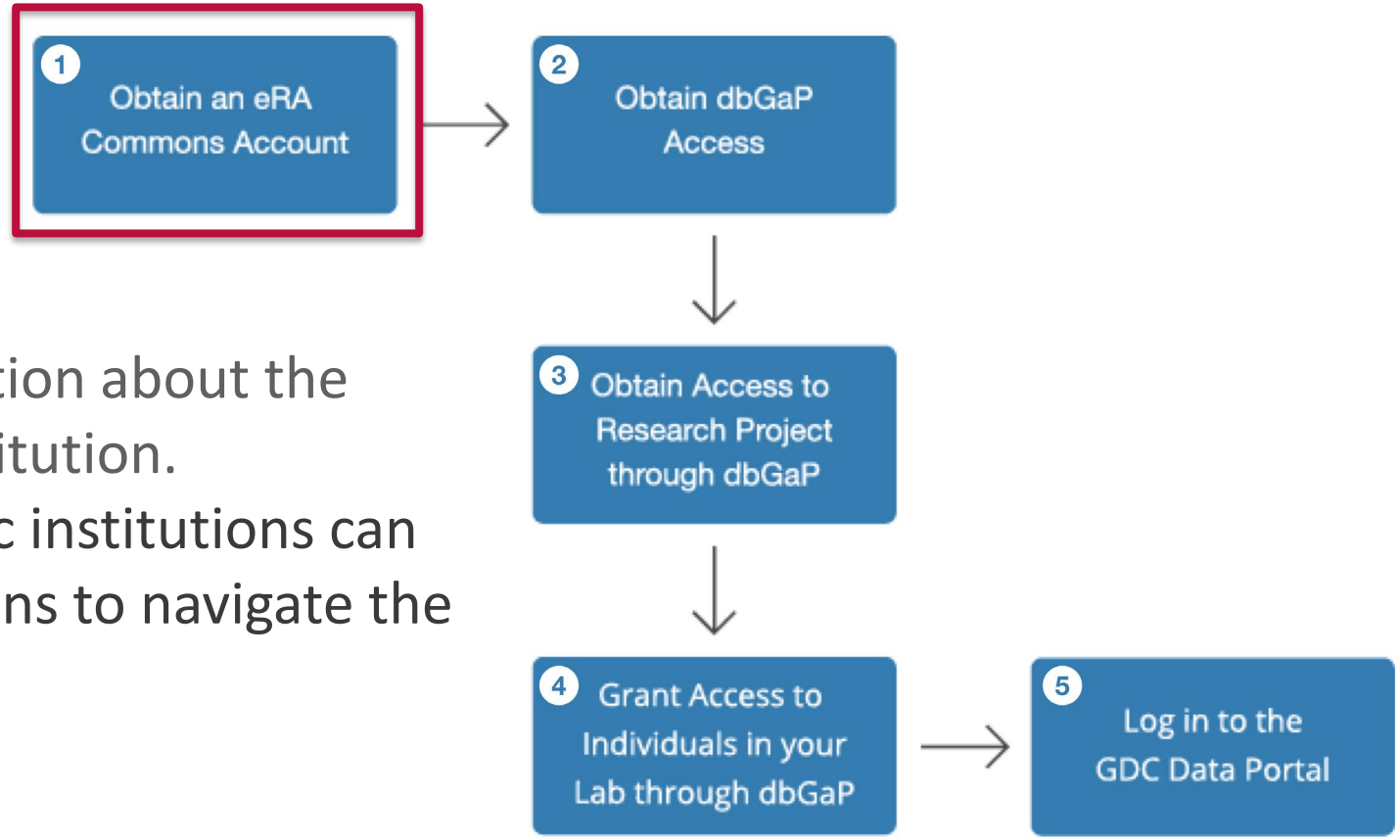
Controlled Access Data

Controlled Data in the GDC

Files in the GDC are either open or controlled-access.
Controlled files usually contain identifiable information, such as genomic sequence
Open files may contain derived and much less identifiable information.

Exp. Strategy	Open	Controlled
Whole Exome	• Masked somatic mutation (MAF)	• Alignments (BAM) • Unfiltered variants (MAF/VCF)
Whole Genome	• Copy number variation (TXT/TSV)	• Alignments (BAM) • SSM variants (VCF) • Structural variants (VCF/BEDPE)
RNA-Seq	• Gene expression (TSV)	• Alignments (BAM) • Splice junction (TSV) • Transcript fusion (BEDPE)
Methylation	All	N/A
Clinical / Biospecimen	All	N/A

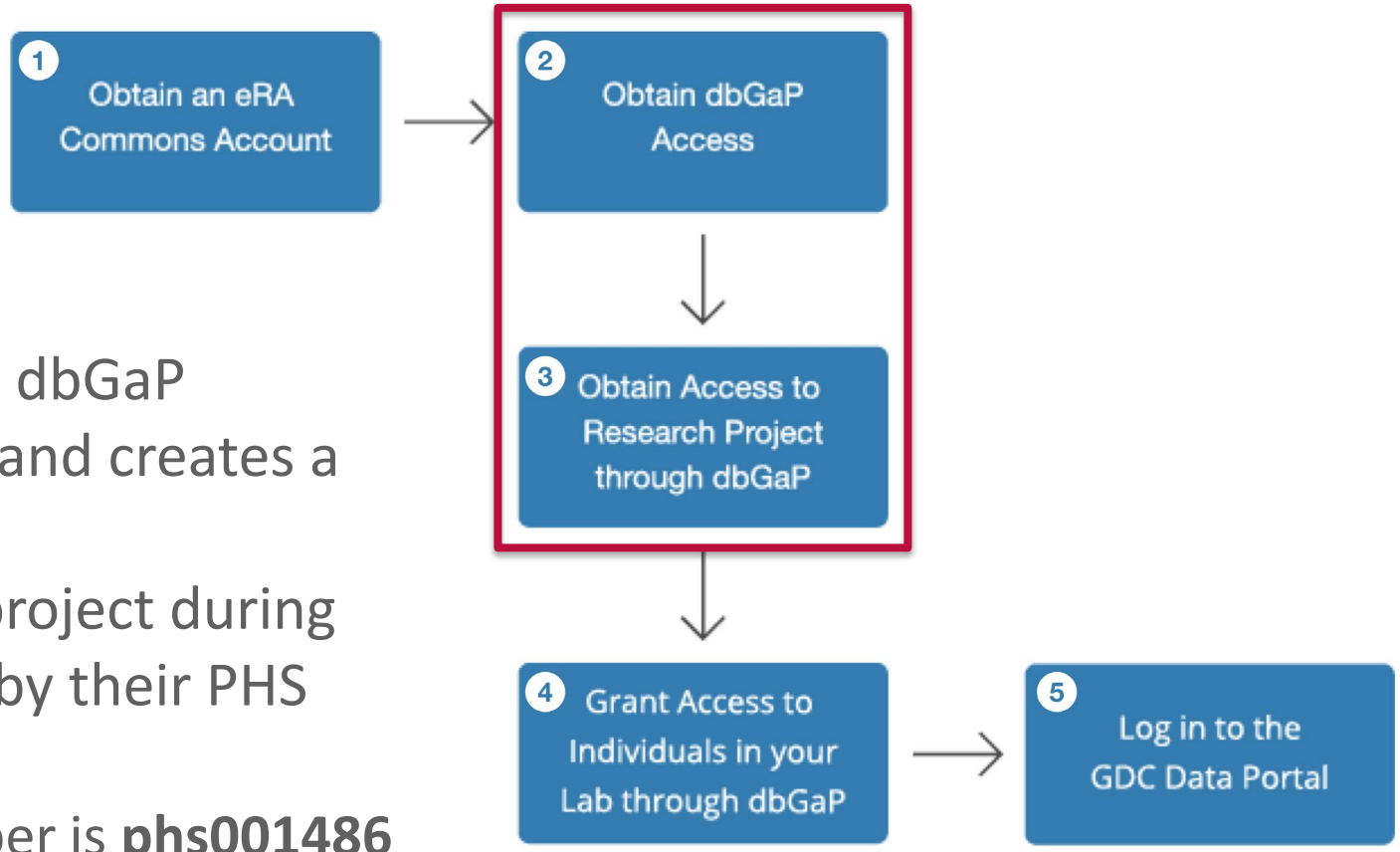
Obtaining Access to Controlled Data in the GDC



eRA Commons ID requires information about the users, their research, and their institution.

Researchers working with academic institutions can often get help from those institutions to navigate the eRA application process.

Obtaining Access to Controlled Data in the GDC

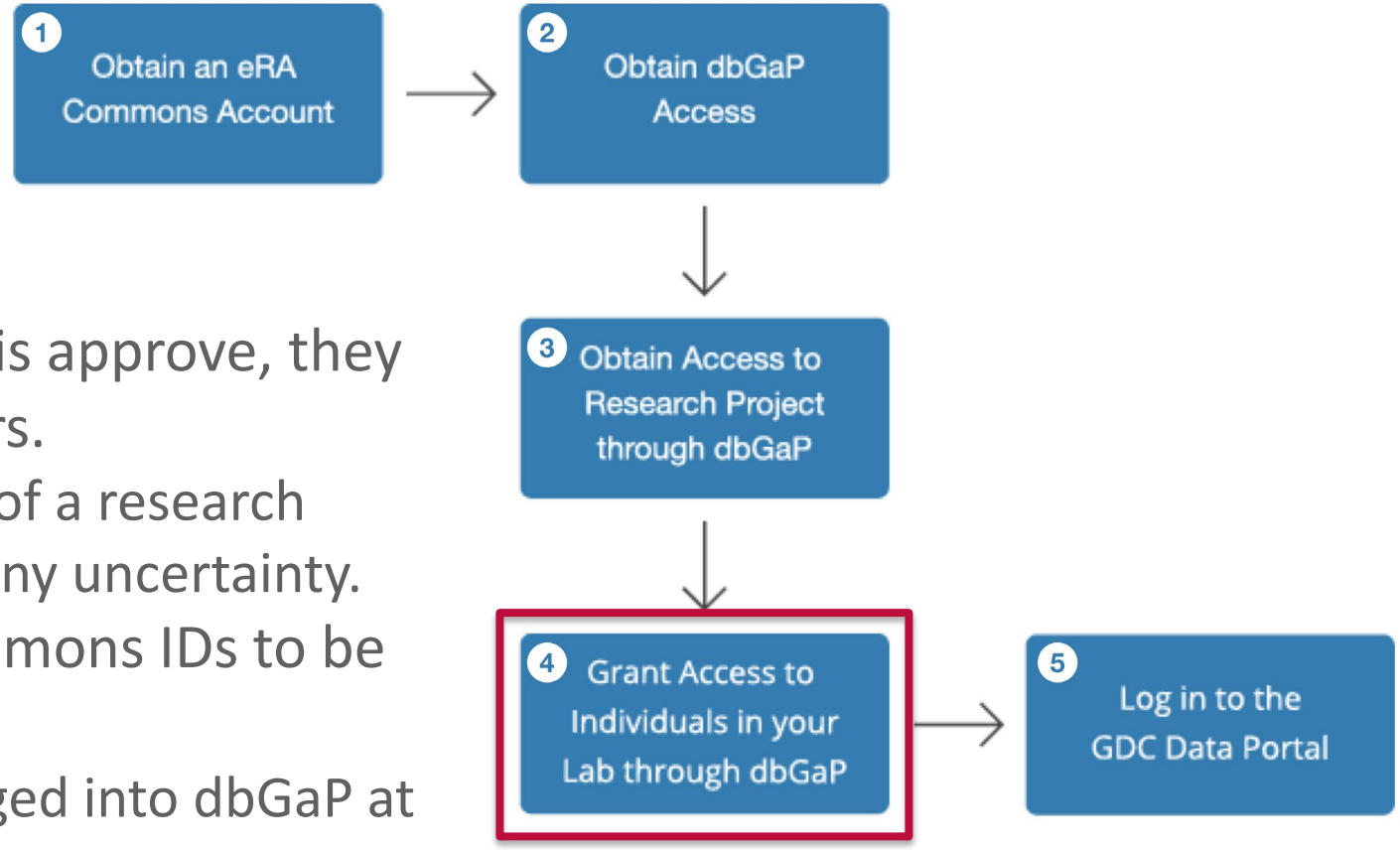


eRA Commons ID is used to log in to dbGaP
The PI of the group logs into dbGaP and creates a dbGaP project.
Specific data sets are added to the project during registration. Data sets are specified by their PHS number.

For the HCMI data set, this number is **phs001486**

Each data set has its own policies, which are outlined when the data set is added to the dbGaP project.

Obtaining Access to Controlled Data in the GDC



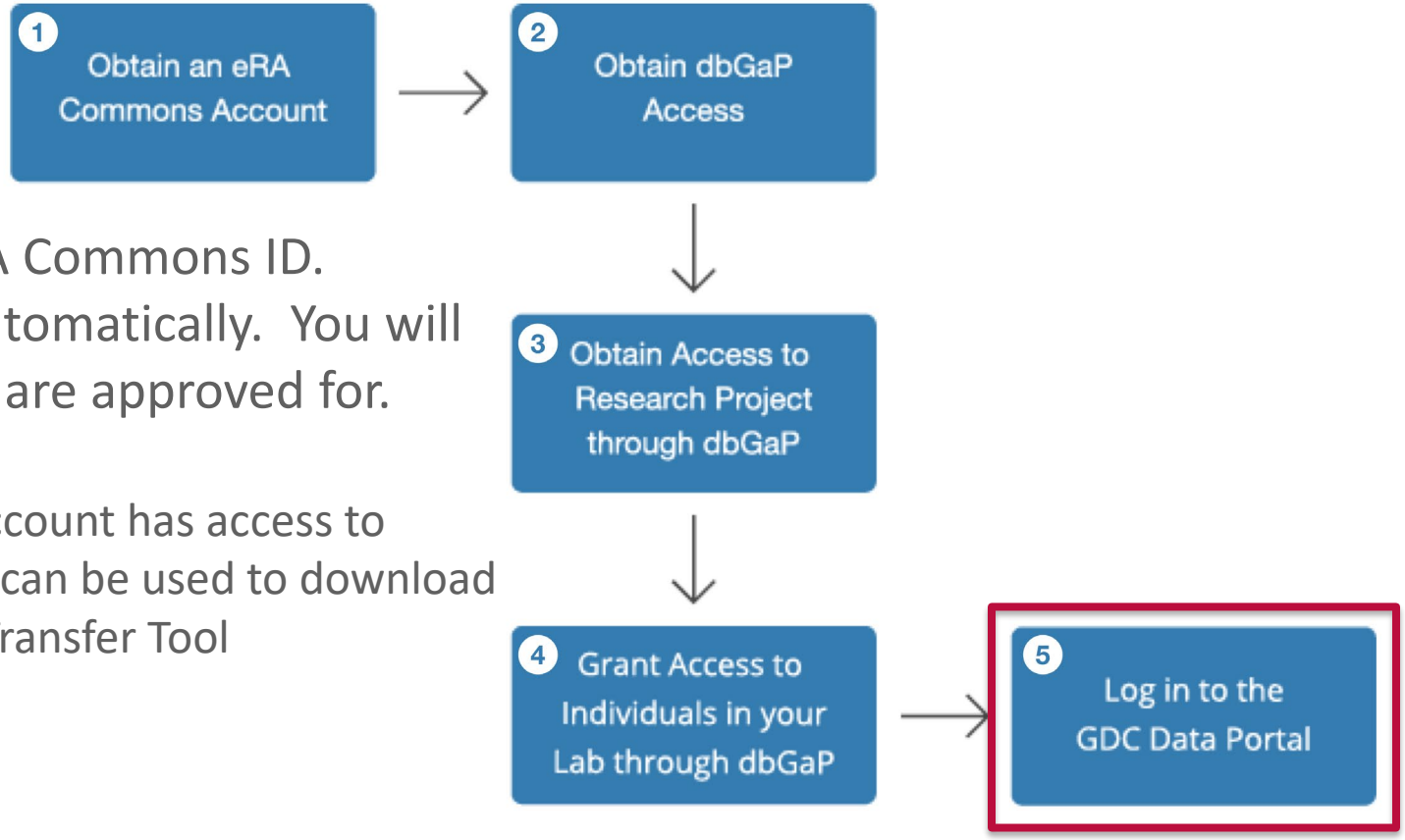
Once the PI's access to the data set is approve, they can provide access to team members.

- Usually this applies to members of a research laboratory, contact dbGaP with any uncertainty.

Team members must have eRA Commons IDs to be granted access.

- Additionally, they must have logged into dbGaP at least once before.

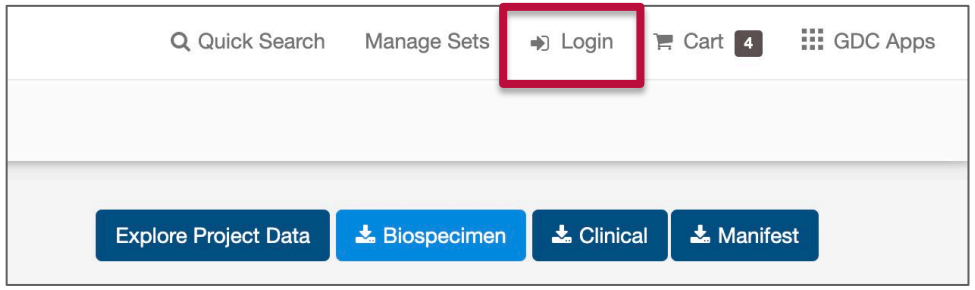
Obtaining Access to Controlled Data in the GDC

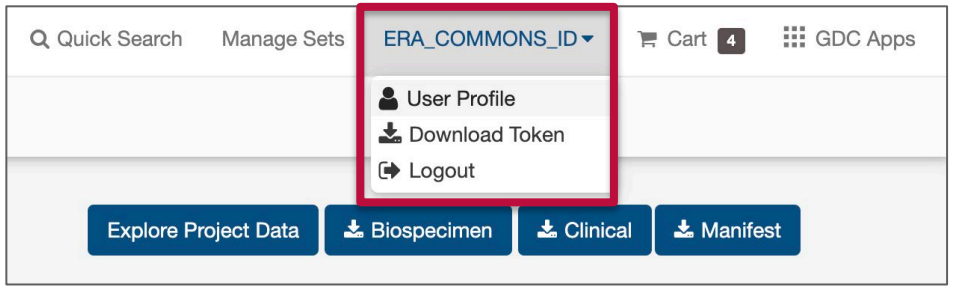


Log into the GDC Data Portal using the eRA Commons ID. Permissions from dbGaP will be applied automatically. You will have access to any controlled data set you are approved for.

Other Options:

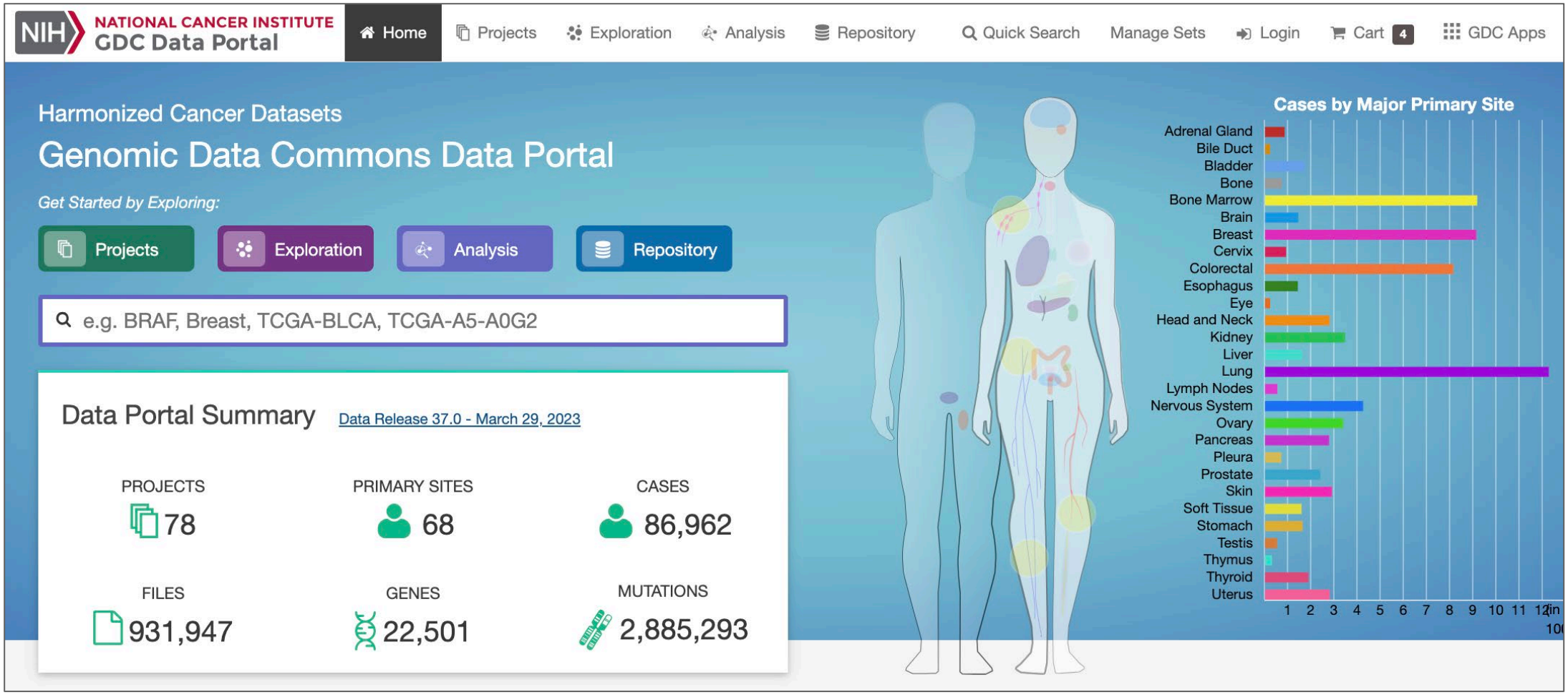
- User Profile: Shows which projects your account has access to
- Download Token: Downloads a token that can be used to download controlled GDC data through API or Data Transfer Tool





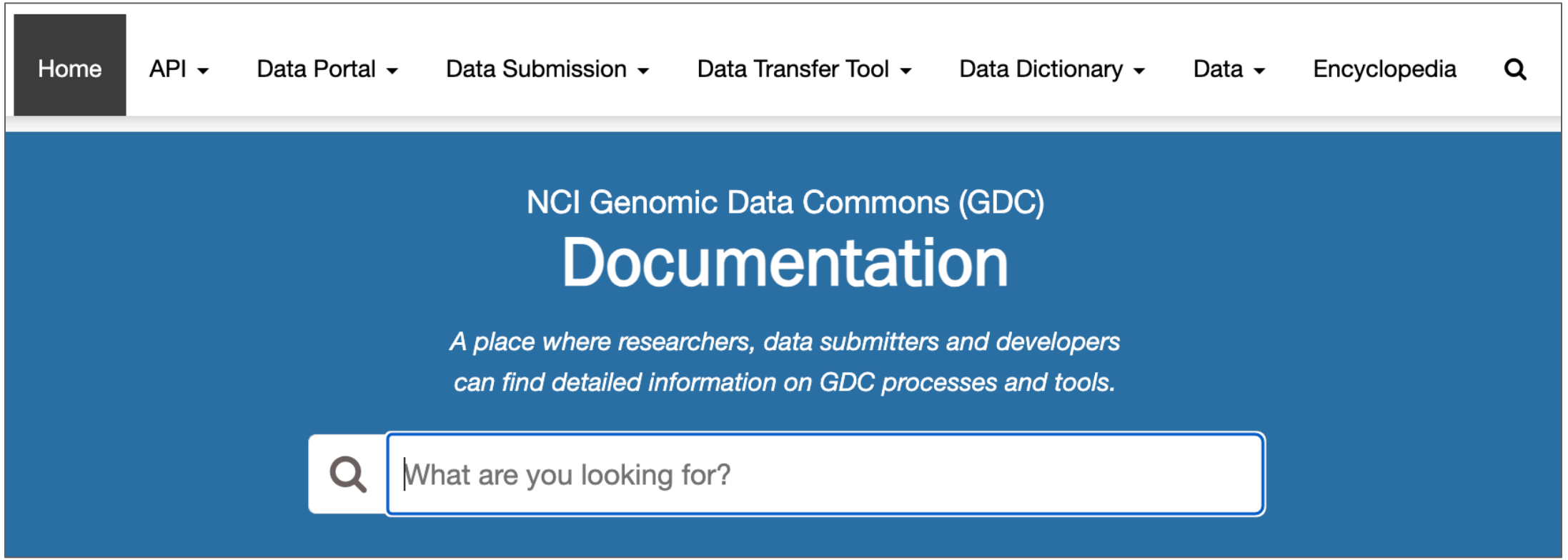
Useful Links

GDC Portal – https://portal.gdc.cancer.gov



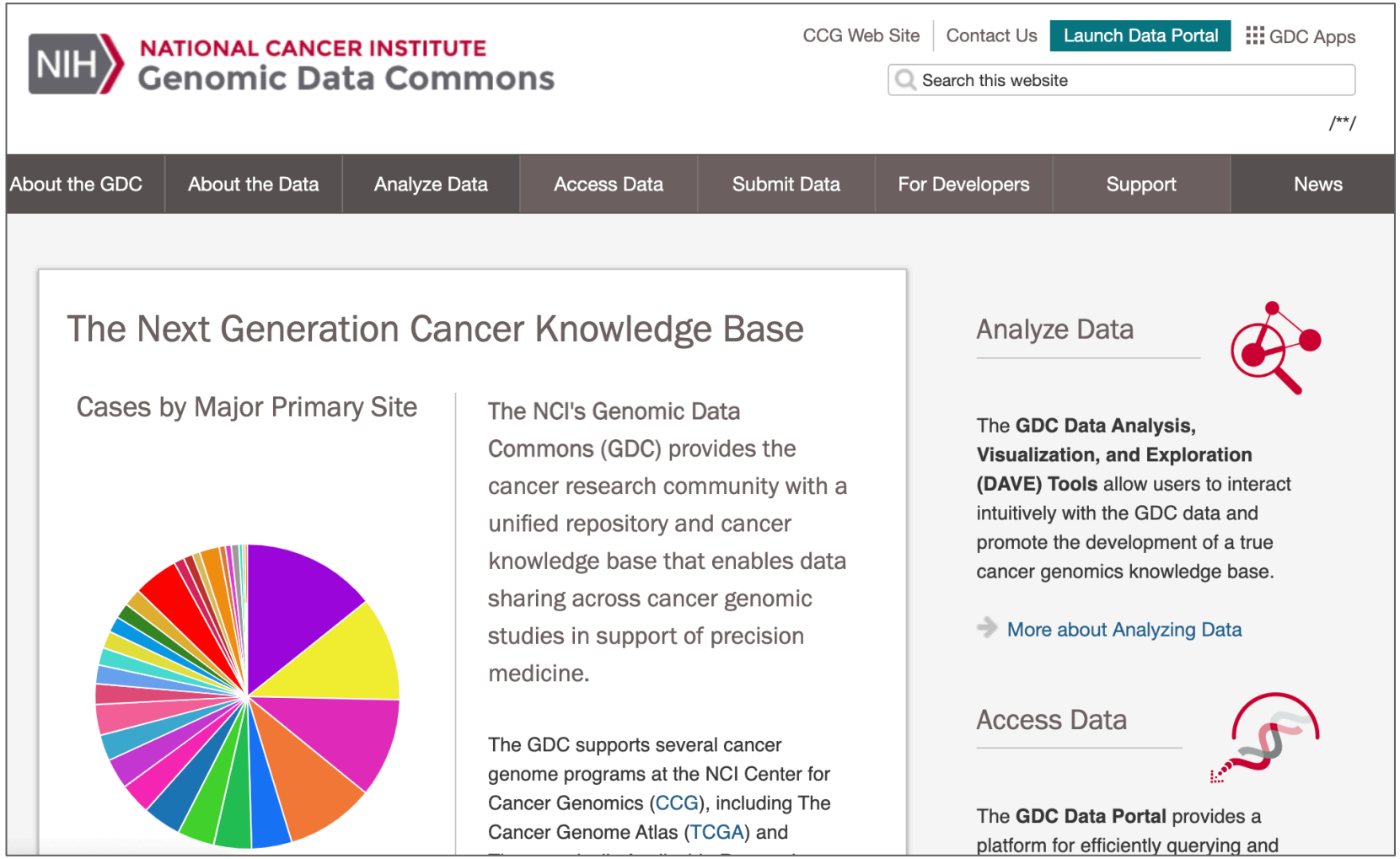
Useful Links

GDC Documentation – https://docs.gdc.cancer.gov



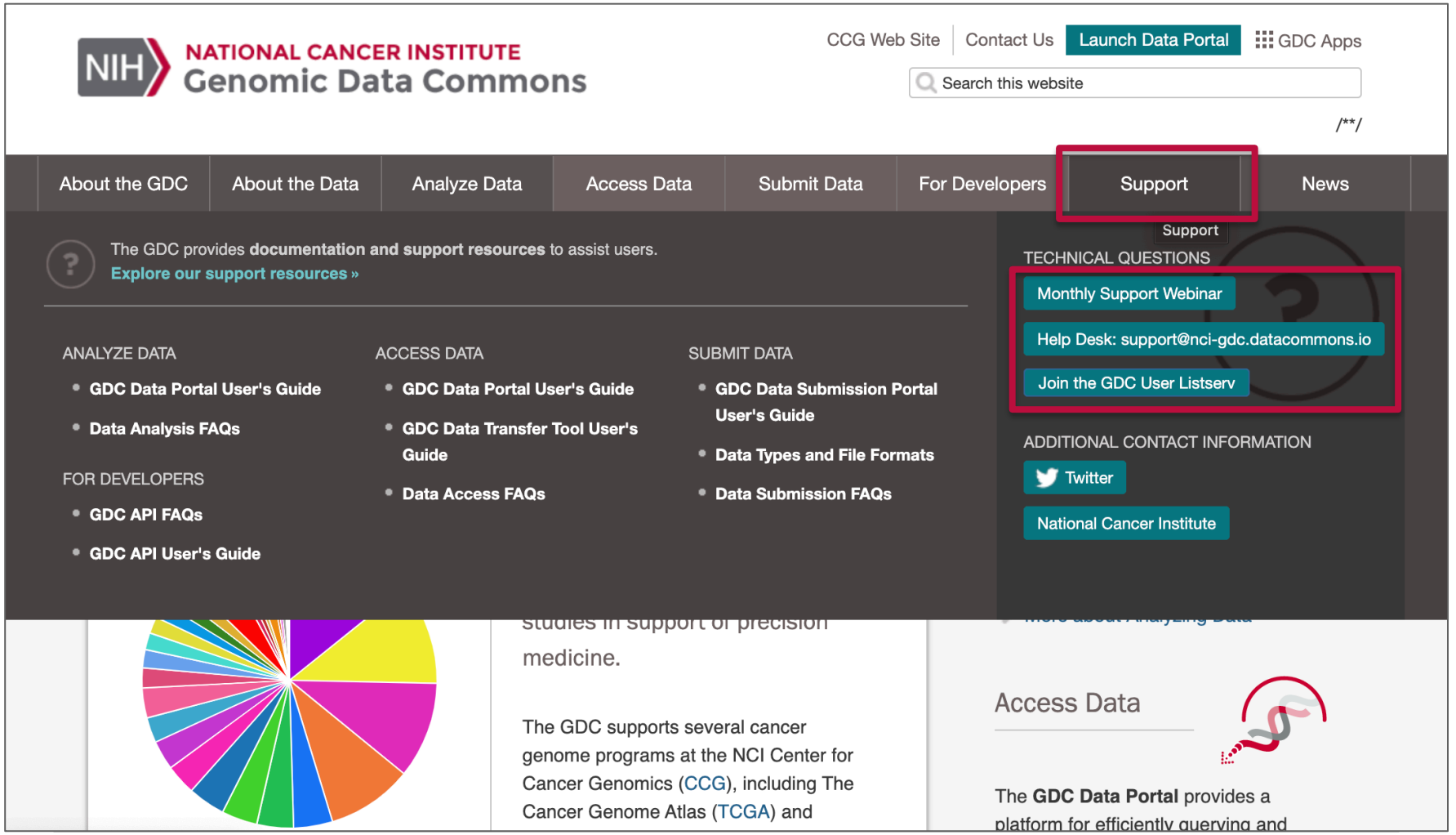
Useful Links

GDC Website – https://gdc.cancer.gov



Useful Links

GDC Help Desk – support@nci-gdc.datacommons.io



U.S. Department of Health & Human Services
National Institutes of Health | National Cancer Institute

https://www.cancer.gov/

1-800-4-CANCER

Produced July 2023

Help us prioritize future HCMI model releases

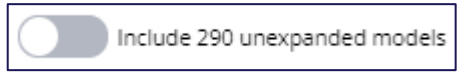
Hundreds of models from over a dozen different tissues have not yet entered our release pipeline:

- Bladder
- Blood and lymphatic system
- Brain
- Breast
- Connective and soft tissue
- Esophagus
- Gynecological tissue
- Head and neck
- Kidney
- Liver
- Digestive tract
- Lung
- Pancreas
- Skin
- Stomach
- Thyroid

1. Browse and search these unreleased models at ATCC

- Use the “Submit your Input” button on the HCMI Landing page
  - Direct link: www.atcc.org/hcmi-input

2. Explore them at the HCMI Searchable Catalog

- Look for the “unexpanded model” toggle near the top → 

3. Email us which models are most relevant for your research

- Contact us at: hcmi@atcc.org