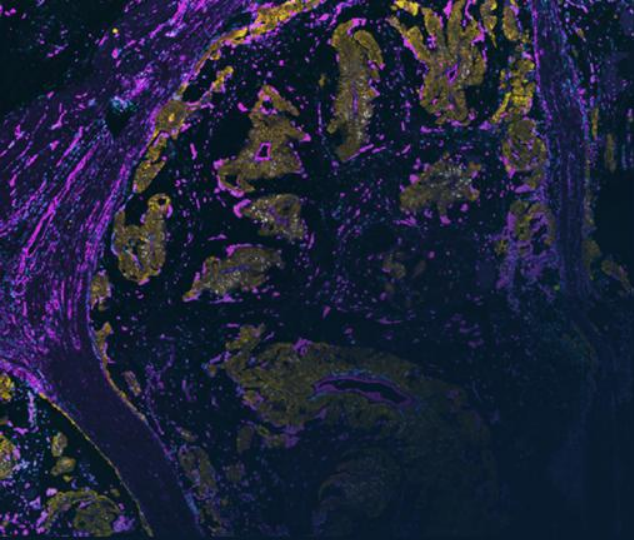
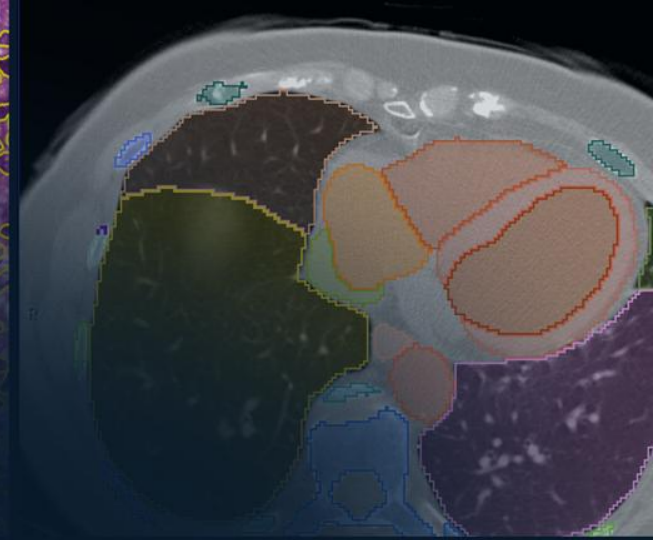




NCI IMAGING DATA COMMONS

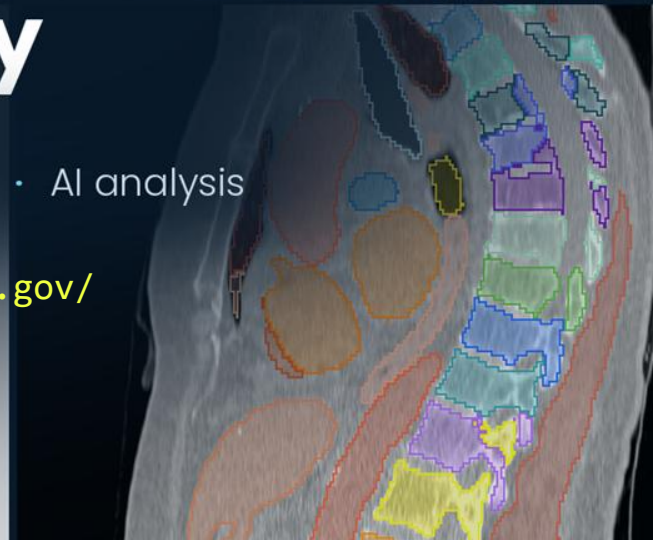
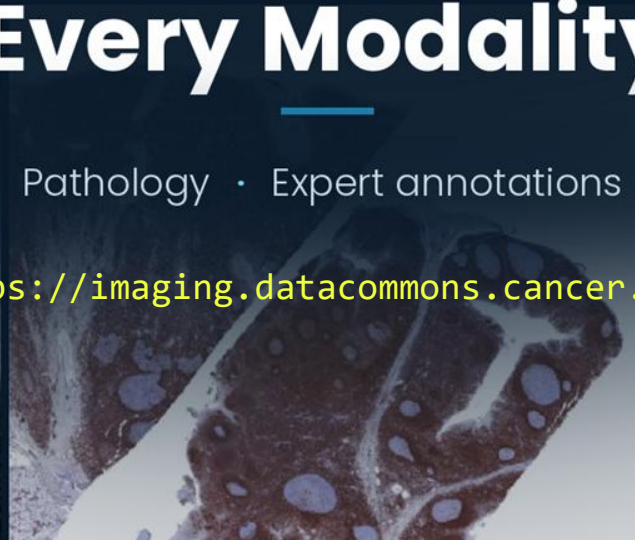


One Cloud, Every Modality



Radiology · Pathology · Expert annotations · AI analysis

<https://imaging.datacommons.cancer.gov/>



Disclosures / disclaimer

- IDC team has been funded in whole or in part with Federal funds from the NCI, NIH, under task order no. HHSN26110071 under contract no. HHSN261201500003I.
- Free cloud hosting and out of cloud egress of IDC data are enabled through the partnership with the respective public datasets programs of Google and Amazon Web Services. IDC team has no financial support from commercial entities.
- We are grateful to the Advanced Cyberinfrastructure Coordination Ecosystem (ACCESS) for the JetStream2 credits allocation.
- The views presented do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Outline

- IDC viewers - what they are and how to use them
- Navigating IDC content outside GDC - IDC agent and human interfaces
- We need your feedback!

👤 Changed 9 hours ago

A guided tour of the NCI Imaging Data Commons viewers

18 samples chosen to show what the IDC viewers can do. Each opens in a browser with one click and displays something the others do not. Ordered so the tables can be walked top to bottom as a demonstration script.

IDC data version	v24 (released 2026-04-16)
Metadata source	IDC hosted service — MCP server api.imaging.datacommons.cancer.gov/mcp and the REST API /v3
GDC linkage source	gdc_idc_mapping.parquet , IDC release artifacts (2026-07-11)
Compiled	2026-08-28

Everything here renders in the viewer. IDC also stores objects carrying no displayable image — dose grids, registrations, surface meshes, clinical data encoded as structured reports — and those are deliberately excluded, since a link that opens to an empty panel undercuts a demonstration.

How to read these tables

- **The Source column carries everything about provenance.** Line one is the collection and patient ID, hyperlinked — clicking it opens that exact series in the viewer. Line two is the license and, where one exists, the matched GDC case.

Open access:
<https://hackmd.io/G2FtDuYRLe1KByLDhz4uA?view>

Patient ^

ID: TCGA-AF-2687
Name: TCGA-AF-2687
Sex:
Birthdate:
Age:

Study ^

Accession #: TCGA-AF-2687
ID: TCGA-AF-2687
Date: 2011-08-03
Time: 21:30:08

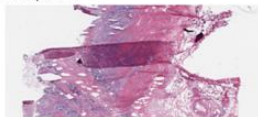
Clinical Trial ^

Sponsor Name: TCGA
Protocol ID: TCGA-READ
Protocol Name: TCGA Rectum
adenocarcinoma
Site Name: Christiana Healthcare

Slides ^

TCGA-AF-2687-01A-02-BS2

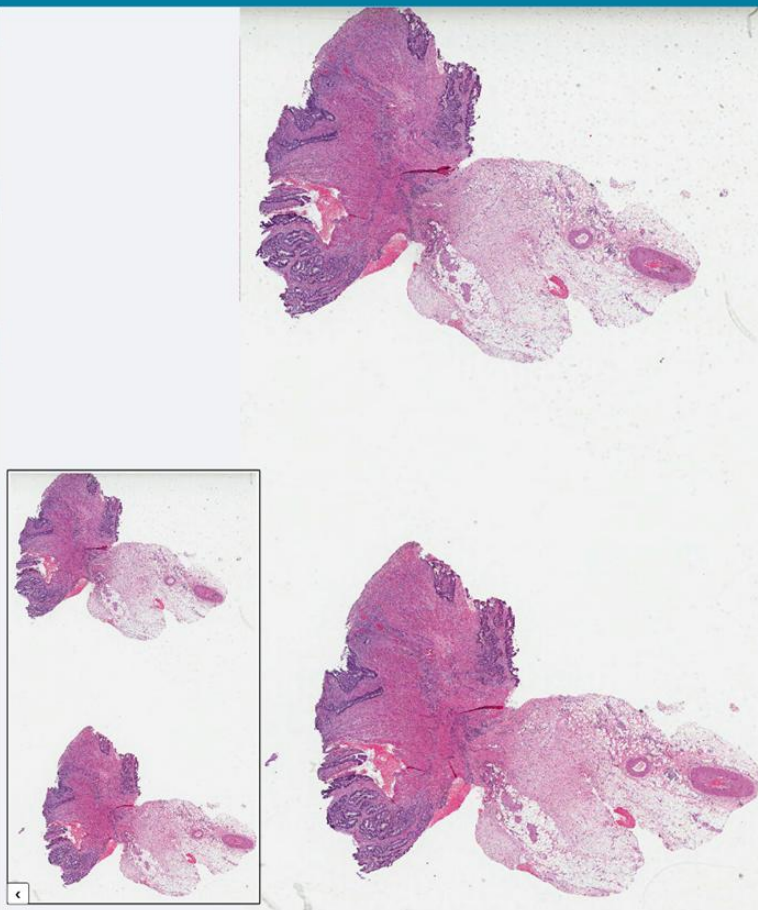
Description: —



Frozen HE TP BS2

TCGA-AF-2687-01A-01-BS1

Description: —



(-13.81480, -8.96940)

Specimens ^

TCGA-AF-2687-01Z-00-DX1

Description:

FFPE HE TP DX1

Anatomical structure:

Rectum

Primary Anatomic Structure Modifier:

Neoplasm, Primary

Collection method:

Removal

Parent specimen:

TCGA-AF-2687-01

Parent specimen:

TCGA-AF-2687-01Z

Parent specimen:

TCGA-AF-2687-01Z-00

Tissue fixative:

Formalin

Tissue embedding medium:

Paraffin wax

Equipment ^

Manufacturer:

Leica Biosystems

Model Name:

**Aperio converted by
com.pixelmed.convert.TIFFToDicom**

Device Serial Number:

SS1302

Software Versions:

**v10.2.41Sat May 28 11:42:06 EDT
2022**

Institution Name:

Christiana Healthcare

Memory: 76.95 MB of 4.09 GB (1.8%) Remaining: 4.02 GB

Patient ^

ID: TCGA-AF-2687
Name: TCGA-AF-2687
Sex:
Birthdate:
Age:

Study ^

Accession #: TCGA-AF-2687
ID: TCGA-AF-2687
Date: 2011-08-03
Time: 21:30:08

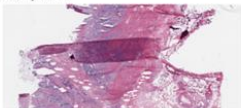
Clinical Trial ^

Sponsor Name: TCGA
Protocol ID: TCGA-READ
Protocol Name: TCGA Rectum
adenocarcinoma
Site Name: Christiana Healthcare

Slides ^

TCGA-AF-2687-01A-02-BS2

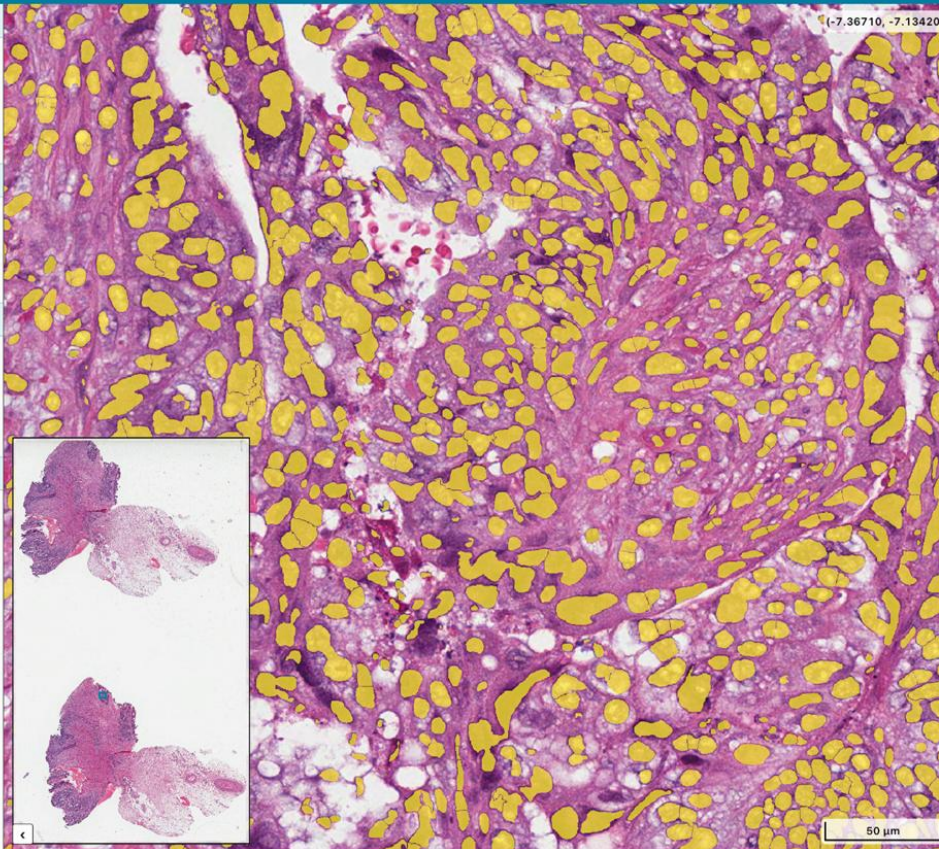
Description: —



Frozen HE TP BS2

TCGA-AF-2687-01A-01-BS1

Description: —



Algorithm Name:
Stony Brook TIL Segmentation CNN 2018
Algorithm Type:
AUTOMATIC
Segmentation Type:
BINARY



TILs Present



Property Type:
Tumor infiltration by lymphocytes present
Property Category:
Morphologically abnormal structure
Algorithm Name:
Stony Brook TIL Segmentation CNN 2018
Algorithm Type:
AUTOMATIC
Segmentation Type:
BINARY



Nuclei



Property Type:
Nucleus
Property Category:
Anatomical Structure
Algorithm Name:
Pan-Cancer-Nuclei-Seg
Algorithm Type:
AUTOMATIC
Segmentation Type:
BINARY



TILs Present

Memory: 148.85 MB of 4.09 GB (3.6%) Remaining: 3.95 GB

Patient

ID: TCGA-AF-2687
Name: TCGA-AF-2687
Sex:
Birthdate:
Age:

Study

Accession #: TCGA-AF-2687
ID: TCGA-AF-2687
Date: 2011-08-03
Time: 21:30:08

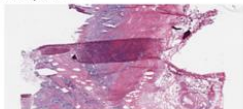
Clinical Trial

Sponsor Name: TCGA
Protocol ID: TCGA-READ
Protocol Name: TCGA Rectum
adenocarcinoma
Site Name: Christiana Healthcare

Slides

TCGA-AF-2687-01A-02-BS2

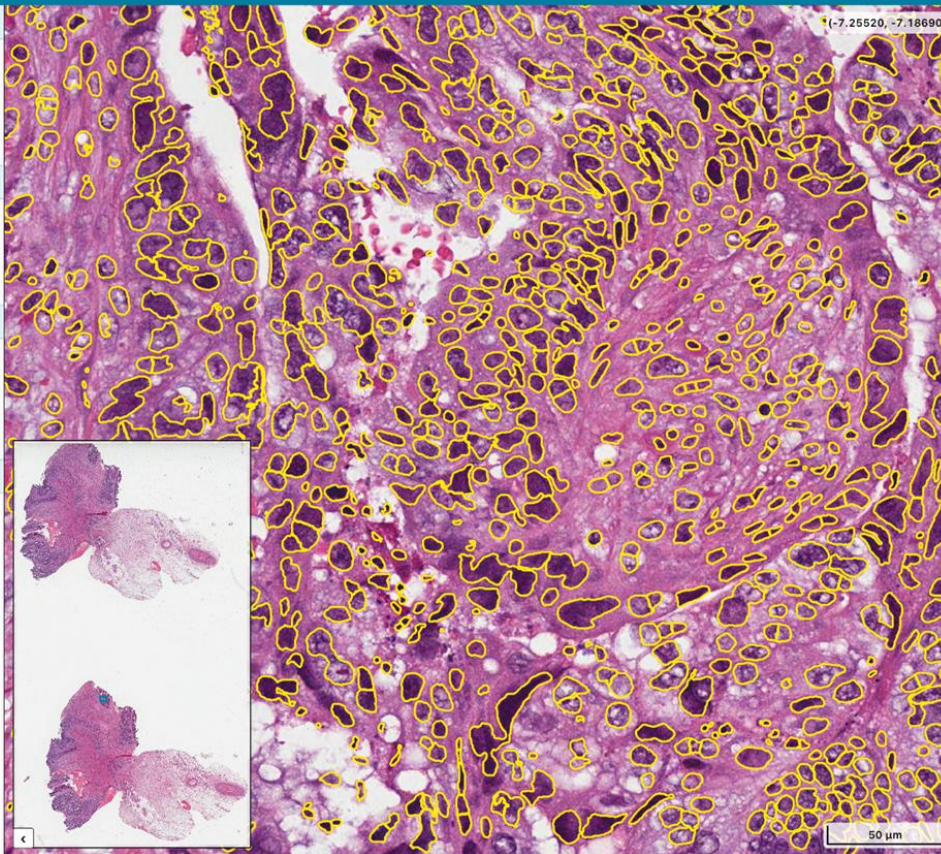
Description: —



Frozen HE TP BS2

TCGA-AF-2687-01A-01-BS1

Description: —



Model Name:

Aperio converted by
com.pixelmed.convert.TIFFToDicom

Device Serial Number:

SS1302

Software Versions:

v10.2.41Sat May 28 11:42:06 EDT
2022

Institution Name:

Christiana Healthcare

Optical Paths



1



Tissue stain:

hematoxylin stain

Tissue stain:

water soluble eosin stain

Annotations

Annotation Groups

All



Nuclei



Property type:

Nucleus

Property category:

Anatomical Structure

Graphic type:

POLYGON

Annotation coordinate type:

2D

Segmentations

Memory: 251.43 MB | of 4.09 GB | (6.0%) | Remaining: 3.85 GB

BCL2 T

HTMCP-01-06-00611-01-Ki67

Description: —



Ki67 T

HTMCP-01-06-00611-01-CD20

Description: —



CD20 T

HTMCP-01-06-00611-01-HE

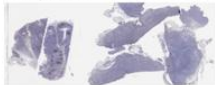
Description: —



HE T

HTMCP-01-06-00611-01-CD10

Description: —



CD10 T

HTMCP-01-06-00611-01-BCL6

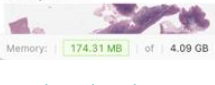
Description: —



BCL6 T

HTMCP-01-06-00611-01-INT

Description: —



Memory: 174.31 MB of 4.09 GB (4.2%) Remaining: 3.92 GB

(16.93430, -57.95550)

Specimens

HTMCP-01-06-00611-01Z

Description:

BCL2 T

Anatomical structure:

Lymph node

Primary Anatomic Structure Modifier:

Tumor

Equipment

Manufacturer:

Leica Biosystems

Model Name:

Aperio converted by com.pixelmed.convert.TIFFToDicom

Device Serial Number:

SS7168CNTLR

Software Versions:

v12.0.15Sun Dec 21 14:47:09 EST

2025

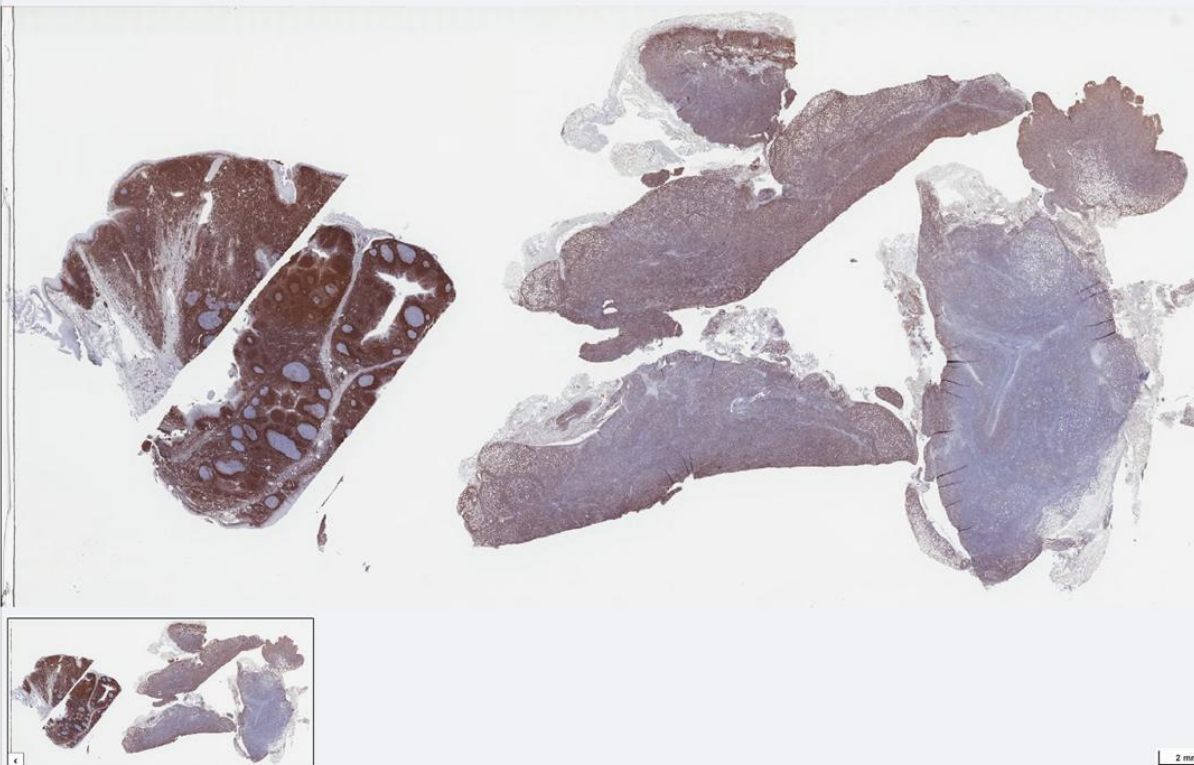
Optical Paths

1

Tissue stain:

BCL2 Ag Immune stain

Annotations



2 mm

Patient ^

ID: HTA7_926
Name: HTA7_926
Sex: Male
Birthdate:
Age:

Study ^

Accession #: HTA7_926
ID: HTA7_926
Date: 1970-01-01
Time: 00:00:00

Clinical Trial ^

Sponsor Name: NCI
Protocol ID: HTAN-HMS
Protocol Name:
Site Name: Harvard Medical School
Time Point ID: Initial Diagnosis

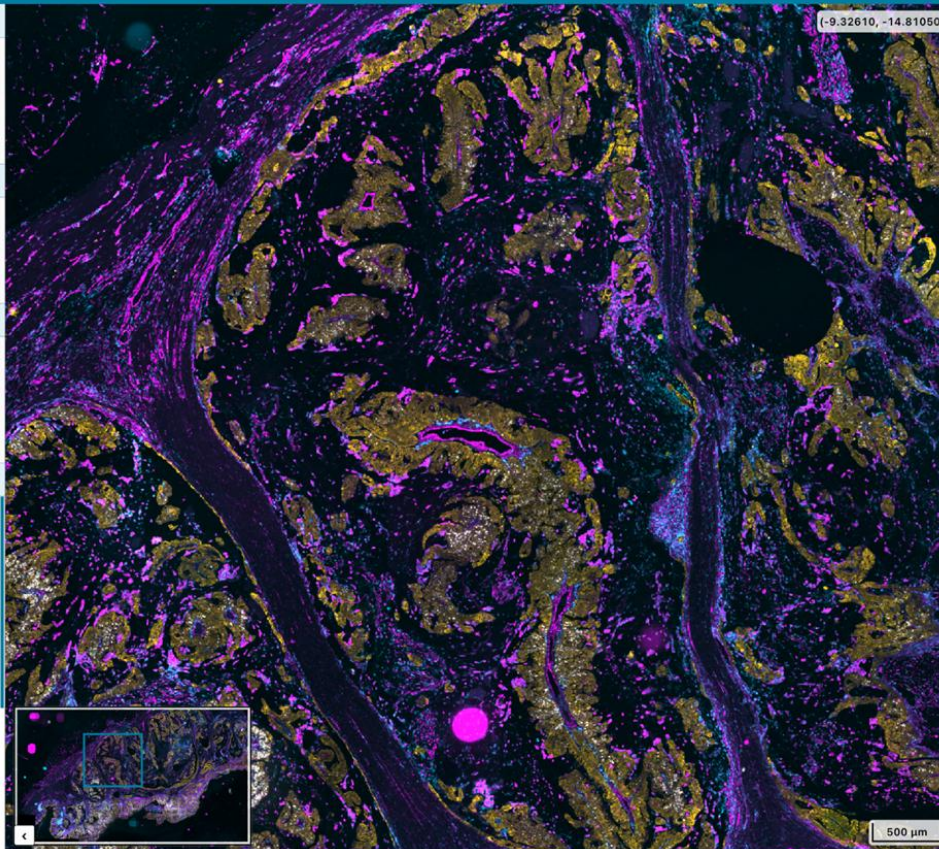
Slides ^

HTA7_926_1000

Description: —

SM

t-CyCIF



(-9.32610, -14.81050)

Specimens ^

HTA7_926_1000

Description:
t-CyCIF
Anatomical structure:
Rectum
Primary Anatomic Structure Modifier:
Neoplasm, Primary
Collection method:
Resection
Parent specimen:
HTA7_926_2

Equipment ^

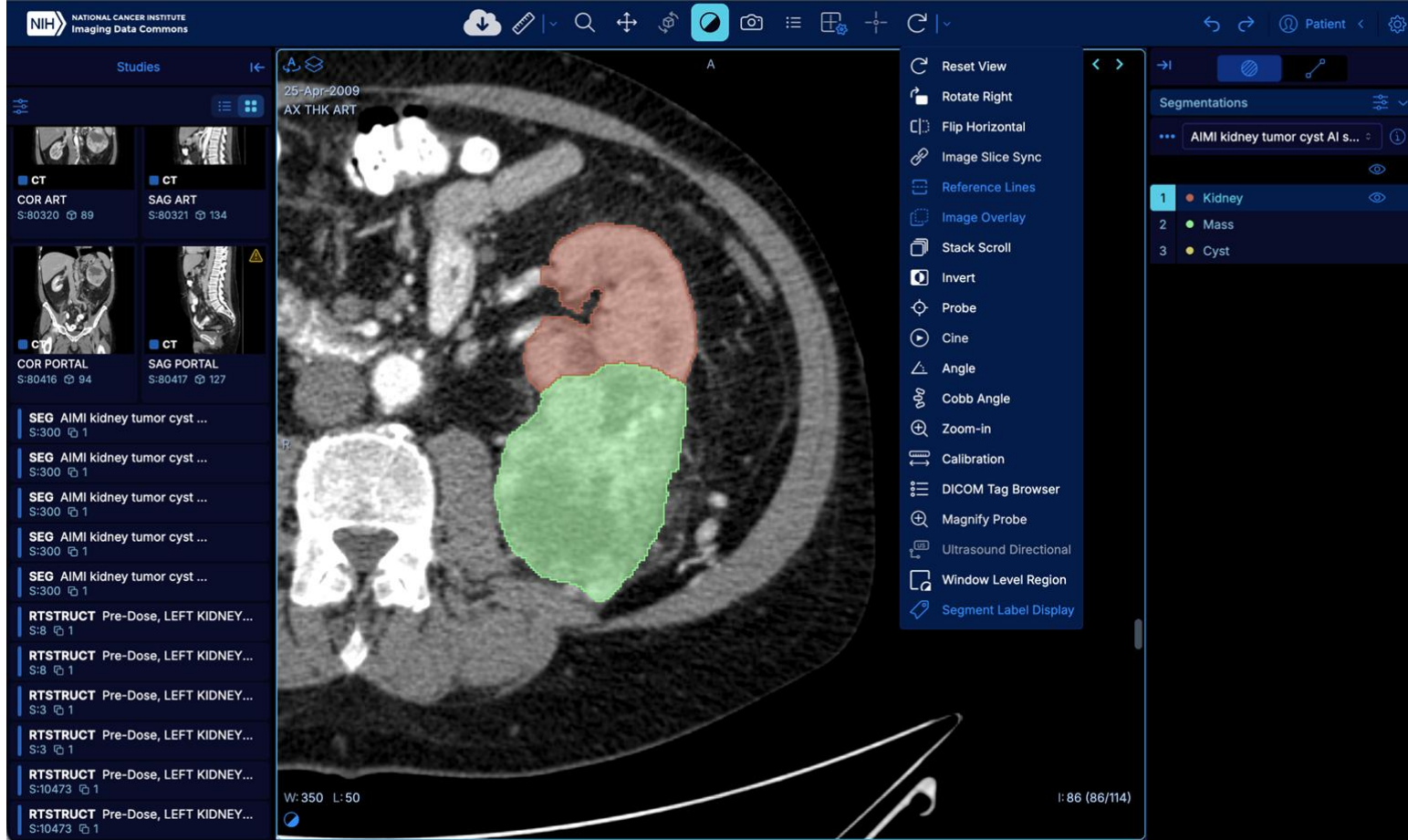
Manufacturer:
RareCyte
Model Name:
CyteFinder converted by
com.pixelmed.convert.TIFFToDicom
Device Serial Number:
**3b10d82dc47dd3a2:2f641cf8:1818bc
33985:-7c26**
Software Versions:
Sat Jun 18 09:29:28 EDT 2022
Institution Name:
Harvard Medical School

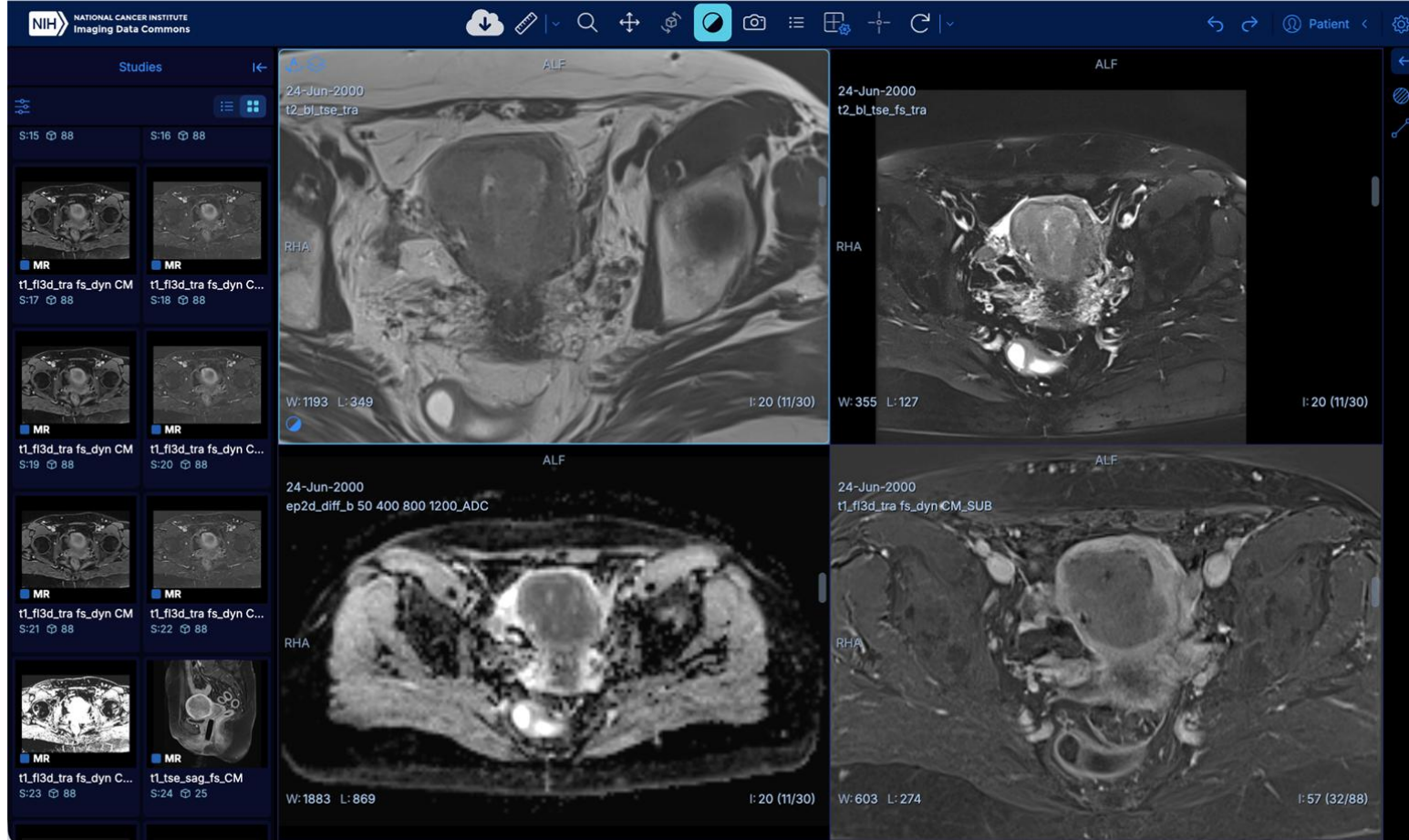
Optical Paths ^

☒ 10: Pan-cytokeratin

Illumination wavelength:
555 nm
 Illumination color:
Narrow
Tissue stain:

Memory: | 298.05 MB | of | 4.09 GB | (7.1%) | Remaining: | 3.80 GB





257/512

We need your feedback about IDC visualization!

- What features are you missing from the original GDC viewer?
- What new features would you like to see?
- What can we make better?

Work in progress - planned for release Fall 2026:

- Speed up initial loading - LIVE
- Simplify viewer image download interface

Welcome!

 Getting started

Core functions

Frequently Asked Questions

Support

Key pointers

Publications

IDC team

Acknowledgments

Jobs

DATA

Introduction

Data model

Licensing and attribution

Organization of data

Data versioning

Downloading data

Data release notes

Data known issues

AI ASSISTANTS[Start here](#)

Hosted server (MCP)

IDC agent skill

REST API

Powered by GitBook

AI assistants

Copy

Start here

You can ask an AI assistant to search IDC, size and refine a cohort, check licenses, generate citations, and hand you a download command — in plain conversation, without writing any code.

There are two ways to give an assistant that ability, and a third option if you would rather write the code yourself.

Which one should I use?

Your situation	Use
You want the fastest start, or your assistant cannot run code	Hosted server (MCP) — paste one URL into your client. Nothing to install, no account, no API key.
You want the widest coverage, and your assistant can run Python (Claude Code, Cursor, Codex, Gemini CLI; claude.ai or Claude Desktop with code execution enabled)	IDC agent skill — adds BigQuery, DICOMweb, direct bucket access, and the digital-pathology and clinical-data guides that the hosted server does not expose.
You are writing the script, application, or notebook yourself	REST API — plain HTTP/JSON, in any language.

These are not mutually exclusive. A good default is to start with the hosted server, then add the skill when you need the extra reach — the skill's own setup guide recommends running both.

 A common question we hear from the users: "Why can't I just use vanilla LLM? It seems to know about IDC already!". It is true that it is easiest to use vanilla LLM. The problem is that without IDC MCP or skill, it is very likely to ground its answer in web search, which will often be imprecise, incomplete, or plain wrong. The help you understand the difference, we prepared the following slides.

IMAGING DATA CYTOMIX - ENABLEMENT GUIDE
Prepared: Summer 2026

Three ways to use your AI assistant

ON THIS PAGE

[Which one should I use?](#)

What you can ask for

What all three paths share

Precautions

In this section

Was this helpful?

Published for free with GitBook's
Community Plan

Sponsored via GitBook

Are there any prostate cancer images in IDC?

<https://claude.ai/share/f2ce1aae-a27d-4d00-9016-a91af20528a7>

Used IDC MCP - prod integration, viewed a file, loaded tools

Yes — quite a lot. IDC v24 has 13 prostate-focused collections (~2,800 subjects, ~37,000 series), plus several mixed-cancer collections that include prostate cases.

Prostate-focused collections

Collection	Subjects	Series	Modalities	Notes
<code>prostatex</code> (PROSTATEx)	346	19,177	MR, SEG, SR	mpMRI lesion-classification challenge data; includes non-cancer controls
<code>prostate_mri_us_biopsy</code>	1,151	10,373	MR, US, SEG, M3D	MRI-ultrasound fusion biopsy, with target/gland annotations
<code>tcga_prad</code> (TCGA-PRAD)	500	3,336	CT, MR, PT, SM, SEG, ANN	radiology and pathology slides, plus clinical + genomics
<code>psma_pet_ct_lesions</code>	378	1,791	PT, CT, SEG	PSMA PET/CT with lesion segmentations
<code>cmb_pca</code> (CMB-PCA)	51	372	CT, MR, PT, NM, SM, DX, RF	Cancer Moonshot Biobank, multimodal
<code>prostate_diagnosis</code>	92	368	MR	
<code>qin_prostate_repeatability</code>	15	330	MR, SEG, SR	test-retest mpMRI
<code>prostate_fused_mri_pathology</code>	28	324	MR	MRI fused with histopathology
<code>prostate_anatomical_edge_cases</code>	131	262	CT, RTSTRUCT	radiotherapy contours, difficult anatomy
<code>naf_prostate</code>	9	214	PT, CT	NaF bone PET
<code>prostate_mri</code>	26	182	MR	
<code>pelvic_reference_data</code>	58	116	CT	prostate + anal, normal pelvic anatomy
<code>prostate_3t</code>	64	64	MR	

CGCI-HTMCP-DLBCL is a digital pathology collection — whole slide images only, no radiology.

What's in it

43 subjects from the HIV+ Tumor Molecular Characterization Project (Diffuse Large B-Cell Lymphoma), part of the Cancer Genome Characterization Initiative. 496 slides, 2,429 DICOM instances, 545 GB total. Every image is Modality **SM** (slide microscopy), brightfield, scanned at 40× with 0.25 μm pixel spacing. These were originally Aperio SVS files from the GDC, converted to DICOM.

Stains per subject

All 43 subjects have the full IHC panel — BCL2, BCL6, CD10, CD20, CD3, CD79a, Ki-67, MUM1, TP53, and EBER (one slide each, 43 series per marker). 42 subjects have an H&E slide; the remaining one has two serial H&E sections instead (S1-HE, S2-HE). 21 subjects also have an INT stain, and one has CD138. That works out to roughly 11–13 slides per subject.

Tissue sites (recorded per specimen, one site per subject):

Site	Subjects
Hematopoietic system	17
Lymph node	14
Liver	5
Kidney	2
Hematopoietic/reticuloendothelial	2
Pleura, brain, bone	1 each

Scale and structure

Each series is a pyramidal image stored as 4–5 instances (base layer plus downsampled levels and the label/overview). Slides average ~1.1 GB, with the largest at 4.4 GB. Base pixel matrices run large — around 170,000 × 80,000 pixels is typical. Worth planning for if you intend to download rather than stream tiles.

License — all 496 series are CC BY 4.0, so commercial use is fine with attribution. Supporting metadata lives on the collection's Zenodo page: <https://doi.org/10.5281/zenodo.17381412>

You can view any slide in the browser without downloading. For example, the BCL2 slide from subject HTMCP-01-07-00160:

<https://viewer.imaging.datacommons.cancer.gov/slim/studies/2.25.281560758186383123414700507233951493378/series/1.3.6.1.4.1.5962.99.1.2956738134.124650608.1772483296854.4.0>

If you want a download manifest for a particular stain or tissue site, or the clinical tables that go with these subjects, I can pull those next.

Tell me about the images for the CGCI-HTMCP-DLBCL IDC collection

<https://claude.ai/share/10ccd518-2ea6-4189-bdf8-af8bb66b14cd>

What can you tell me about subbject HTMCP-01-06-00611 from IDC?

This is a subject in **CGCI-HTMCP-DLBCL** (IDC data release v24) — DICOM-converted whole slide images from the Cancer Genome Characterization Initiative's HIV+ Tumor Molecular Characterization Project for diffuse large B-cell lymphoma. It's pathology only; there's no radiology for this subject.

What's there

One study, dated 2017-07-03, described as "Histopathology": 12 slide microscopy (SM) series, 60 DICOM instances, about 36 GB. Each series is one digitized glass slide stored as a 5-level pyramid, scanned at 40× with 0.25 μm pixel spacing, brightfield. The images are large — roughly 113k-175k pixels wide by 80k-86k tall at full resolution.

The specimen is annotated as **lymph node**, with an admitting diagnosis code of *Diffuse large B-cell lymphoma, NOS*. Container IDs follow the pattern `HTMCP-01-06-00611-01-<stain>`.

The 12 stains — one H&E, ten IHC/ISH markers, plus an INT stain:

Slide	Stain
HE	hematoxylin + eosin
CD20, CD79a	B-cell markers
CD3	T-cell marker
CD10, BCL6, MUM1	Hans-algorithm cell-of-origin panel
BCL2, TP53, Ki-67	prognostic / proliferation
EBER	Epstein-Barr virus-encoded RNA probe
INT	iodonitrotetrazolium

That full panel makes this one of the more completely stained subjects in the collection — many of the 43 subjects lack the INT slide.

Access and terms

View in the browser (SLIM viewer, no download):

<https://viewer.imaging.datacommons.cancer.gov/slim/studies/2.25.11575476516094555195355863238822004089>

All 12 series are CC BY 4.0, so commercial use is fine with attribution. Cite: Clunie, D., Clifford, W., & Fedorov, A. (2026). *CGCI-HTMCP-DLBCL: DICOM converted whole slide images from the Cancer Genome Characterization Initiative (CGCI) HIV+ Tumor Molecular Characterization Project (HTMCP) - Diffuse Large B-Cell Lymphoma* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.17381412>

Two things worth flagging: IDC carries no clinical table for this collection, so HIV status, staging, treatment, and outcome aren't available here — the Zenodo record and the GDC entry for CGCI-HTMCP-DLBCL are where supporting metadata lives. And at ~36 GB for the subject, you'd want to pull individual slides rather than the whole set if you're downloading. I can get you the S3/GCS URLs or an `idc download` manifest for whichever slides you want.

<https://claude.ai/share/00e72e42-56ee-4490-a836-5ad999791550>

Help me select, download and prepare images from IDC for testing an AI model segmenting kidney tumors.

Viewed 2 files

I'll start by checking what the IDC service offers, then find kidney-tumor data with ground-truth segmentations.

Ran 15 commands, used IDC MCP - prod integration, created 5 files · 8 notes

Numbers look plausible. Now the check that actually matters: does the label sit on the kidney?

Created 2 files, ran 2 commands, viewed a file · 1 note



I built you a working pipeline against IDC v24 and validated it on a 9-case pilot.

What I picked: `c4kc_kits` — the KITS19 challenge set. 210 patients, contrast-enhanced CT, expert human-corrected Kidney + Mass segmentations, CC BY 3.0. It's the only IDC collection with genuine manual kidney tumor ground truth. The collection has 411 CT series but only 210 carry a segmentation, so the manifest resolves each SEG to its referenced source CT and skips the rest — 23.5 GB instead of 45.

Two things worth knowing before you trust the numbers it produces:

- KITS19 is in almost every public kidney model's training data. If your model is fine-tuned from one, this measures reproduction rather than generalisation. For an unseen-domain check, the `bamf_aim_annotations` results cover `tcga_kirc/kipr/kich` and `cptac_ccrc` — but those labels are AI-generated, with only 7 expert-reviewed series across all four, so they're a robustness signal, not ground truth. They do add a `Cyst` class KITS lacks, which is the classic kidney-tumor false positive.
- Kidney and tumor labels don't overlap here — label `1` is parenchyma with the mass carved out. If your model outputs kidney as the whole organ, compare against `label > 0` or you'll see a large penalty that's purely a convention mismatch.

The conversion detail that would have bitten you: DICOM SEG stores only non-empty frames in arbitrary order, so stacking `pixel_array` by index produces a plausible-looking but wrong volume. `highdicom` also refuses these files outright because they don't assert spatial-location preservation, and the usual fix (`ignore_spatial_locations=True`) suppresses the check rather than verifying anything. I map each frame onto a CT slice by `ImagePositionPatient` instead, and the script raises if any frame fails to match. The overlay confirms labels land exactly on the kidneys.

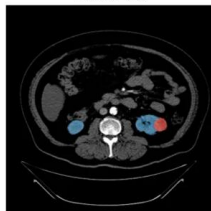
Pilot QC: 9/9 converted, no geometry mismatches, no empty tumors, tumor volumes spanning 1.9–111 mL across 1–5 mm slice thickness and four vendors.

To scale to all 210, change one line in `download_pilot.py` (`pilot = pairs`) and budget ~30 GB disk. Then re-run the QC overlay — volume outliers in `qc_report.csv` are the fastest way to catch a bad case.

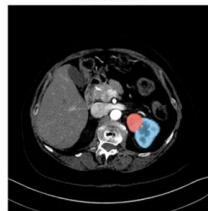
Qc report CSV

PatientID	status	shape	spacing_mm	shapes_match	hu_min	hu_max	kidney_ml	tumor_ml
KITS-00004	ok	512x512x64	0.977x0.977x4.000	True	-1024	2125	439.9	18.8
KITS-00010	ok	512x512x50	0.758x0.758x3.000	True	-1024	1280	352.7	15.0
KITS-00019	ok	512x512x129	0.938x0.938x3.750	True	-2048	1268	530.7	1.9
KITS-00061	ok	512x512x29	0.977x0.977x5.000	True	-2048	1753	595.8	111.0
KITS-00112	ok	512x512x113	0.742x0.742x2.000	True	-1024	1509	456.9	93.8
KITS-00115	ok	512x512x325	0.703x0.703x0.500	True	-1024	3071	275.8	11.1
KITS-00117	ok	512x512x69	0.846x0.846x2.500	True	-3056	1345	623.2	32.2
KITS-00183	ok	512x512x195	0.707x0.707x1.500	True	-1024	1560	313.3	21.9
KITS-00193	ok	512x512x233	0.703x0.703x1.250	True	-2048	1646	261.8	7.9

KITS-00004 z=20



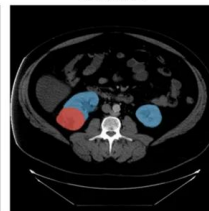
KITS-00010 z=35



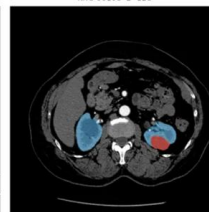
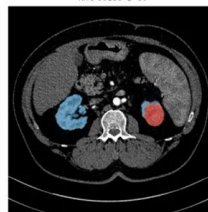
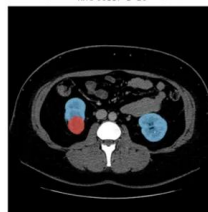
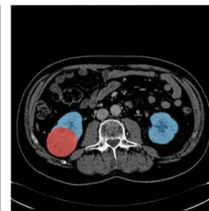
KITS-00019 z=86



KITS-00061 z=9



KITS-00112 z=24



<https://claude.ai/share/9eb40db6-e3fe-473a-ad86-ecb800878cf9>

Readme

Document · MD

download



Andrey Fedorov · Follow

Last edited by [Andrey Fedorov](#) on Sep 8, 2026

IDC imaging coverage for GDC projects

Radiology and pathology imaging held in the NCI Imaging Data Commons, organized by the NCI

Genomic Data Commons project the cases belong to.

- **42 GDC projects** have imaging in IDC, via **52 IDC collections**
- **19** projects have both radiology and pathology, **22** are pathology-only, **1** is radiology-only
- **18,586** IDC patients and **21,342** studies in total
- Built from the IDC↔GDC mapping artifact, the live IDC index (**v24**), and the GDC cases API — see [How this report was generated](#)
- Generated 2026-09-08

Coverage by GDC project

Patient counts are distinct `PatientID` values in IDC. A patient can have both radiology and pathology, so the radiology and pathology columns overlap and do not sum to the total.

GDC project	Coverage	Radiology patients	Pathology patients	Total patients	Total studies	IDC coll
CCDI-MCI	Pathology only	—	4,407	4,407	4,447	ccdi_mc
CDDP_EAGLE-1	Pathology only	—	49	49	49	cddp_ea

https://hackmd.io/@fedorov/SkcGI0u_Mx

*What if I don't have IDC skill
/ MCP configured?*

Your Mileage May Vary (YMMV)

What can you tell me about subject HTMCP-01-06-00611 from IDC?

🔍 Locating a dataset through imaging portals and repositories 30s ▾

Identifying the origin of a specimen code.

Searched the web ▾

- ▶ HTM COMPLIANCE WHITEPAPER www.powercontrol.co.uk
- 📄 HTM 06-01 Electrical Services | PDF | Electrical Substation | Health Care www.scribd.com
- 📄 HTM by Number - NHS Property and Estates Guidance buckshealthcare.nhs.libguides.com
- 📄 HTM 06-01 - Commonly used Generator application ... www.wbpsltd.co.uk
- 📄 Key changes in revised HTM 06-01 outlined www.healthstatejournal.com
- ▶ HTM 06-01: Electrical Services in Healthcare Design | EBSIP electricalbsp.org
- ▶ Why a UPS for healthcare estates should be HTM 06-01 compliant www.powercontrol.co.uk
- 📄 [PDF] HTM 06-01: Electrical services supply - Free Download PDF nanopdf.com
- 📄 Welsh Health Technical Memoranda (WHTMs) & Health Technical Memoranda (...) nwssp.nhs.wales
- 📄 Health Technical Memorandum 06-01 Part A Electrical services supply and distri... www.thenbs.com

Searched the web ▾

- 🌟 HIV+ Tumor Molecular Characterization Project www.nationwidechildrens.org
- 🔗 GitHub - ImagingDataCommons/wg26-2026-connectathon-idc: Toolkit and working ar... github.com
- 🔗 IDC Docs/data/data-release-notes.md at prod · ImagingDataCommons/IDC-Docs github.com
- 📄 Data release notes | IDC User Guide learn.canceridc.dev

Identifying which imaging collections hold the relevant subjects.

Searched the web ▾

Locating a dataset through imaging portals and repositories.

Failed to fetch: [https://api.gdc.cancer.gov/cases?filters=\[{"op":"in","content":{"field":"submitt...](https://api.gdc.cancer.gov/cases?filters=[{) **Failed** ▸

Searched the web ▾

Fetchd: <https://api.gdc.cancer.gov/cases?from=0&size=5&pretty=true> ▸

Diagnosing why the fetch tool returned a cached generic page.

No IDC skill / MCP!
Claude
~4-6 min to response

IDC skill/MCP
configured
Claude
~30 sec to response

Used IDC MCP - prod integration, viewed a file, loaded tools - 1 note

This is a subject in **CGCI-HTMCP-DLBCL** (IDC data release v24) — DICOM-converted whole slide images from the Cancer Genome Characterization Initiative's HIV+ Tumor Molecular Characterization Project for diffuse large B-cell lymphoma. It's pathology only; there's no radiology for this subject.

What's there

One study, dated 2017-07-03, described as "Histopathology": 12 slide microscopy (SM) series, 60 DICOM instances, about 36 GB. Each series is one digitized glass slide stored as a 5-level pyramid, scanned at 40× with 0.25 µm pixel spacing, brightfield. The images are large — roughly 113k-175k pixels wide by 80k-86k tall at full resolution.

The specimen is annotated as **lymph node**, with an admitting diagnosis code of *Diffuse large B-cell lymphoma, NOS*. Container IDs follow the pattern **HTMCP-01-06-00611-01-*<stain>***.

The 12 stains — one H&E, ten IHC/ISH markers, plus an INT stain:

Slide	Stain
HE	hematoxylin + eosin
CD20, CD79a	B-cell markers
CD3	T-cell marker
CD10, BCL6, MUM1	Hans-algorithm cell-of-origin panel
BCL2, TP53, Ki-67	prognostic / proliferation
EBER	Epstein-Barr virus-encoded RNA probe
INT	iodonitroretazolium

That full panel makes this one of the more completely stained subjects in the collection — many of the 43 subjects lack the INT slide.

Access and terms

View in the browser (SLIM viewer, no download):

<https://viewer.imaging.datacommons.cancer.gov/slim/studies/2.25.11575476516094555195355863238822004089>

All 12 series are CC BY 4.0, so commercial use is fine with attribution. Cite: Clunie, D., Clifford, W., & Fedorov, A. (2026). *CGCI-HTMCP-DLBCL: DICOM converted whole slide images from the Cancer Genome Characterization Initiative (CGCI) HIV+ Tumor Molecular Characterization Project (HTMCP) - Diffuse Large B-Cell Lymphoma* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.17381412>

Two things worth flagging: IDC carries no clinical table for this collection, so HIV status, staging, treatment, and outcome aren't available here — the Zenodo record and the GDC entry for CGCI-HTMCP-DLBCL are where supporting metadata lives. And at ~36 GB for the subject, you'd want to pull individual slides rather than the whole set if you're downloading. I can get you the S3/GCS URLs or an **idc download** manifest for whichever slides you want.

No IDC skill / MCP! Claude

What can you tell me about subject HTMC-01-06-00611 from IDC?

Ran 8 commands, searched the web >

HTMC-01-06-00611 is a subject (DICOM PatientID) in the NCI Imaging Data Commons collection **CGCI-HTMC-DLBCL** — the diffuse large B-cell lymphoma arm of the HIV+ Tumor Molecular Characterization Project, part of NCI's Cancer Genome Characterization Initiative. HTMC and the Burkitt Lymphoma Genome Sequencing Project are the two active CGCI projects, and HTMC was a joint effort of the Office of Cancer Genomics and the Office of HIV and AIDS Malignancy aimed at comparing tumors from HIV-positive and HIV-negative patients. [HIV+ Tumor Molecular Cha...](#)

What IDC's index says about this particular subject:

- Male, recorded age 68, one Histopathology study with a study date of 2017-07-03
- 12 whole-slide images, all from tumor tissue, anatomic site lymph node, admitting diagnosis "Diffuse large B-cell lymphoma, NOS"
- Stains: H&E plus a lymphoma IHC/ISH panel — CD3, CD10, CD20, CD79a, BCL2, BCL6, MUM1/IRF4, Ki-67, TP53, EBER (in-situ probe for EBV), and one series labeled INT
- Scanned at 40× objective, ~0.25 µm/pixel, with pixel matrices in the range of roughly 113k-175k by 80k-86k; ~35 GB total across the 12 series
- Public data under CC BY 4.0, hosted in the [idc-open-data](#) bucket, first added in IDC v24

For context, the whole collection is 496 whole slide images from 43 subjects, including H&E and IHC/ISH stains (BCL2, BCL6, CD10, CD20, CD3, CD79a, EBER, Ki-67, MUM1, TP53) for most subjects plus INT and CD138 for subsets, converted from GDC Aperio SVS to DICOM slide microscopy. This subject has the median number of slides (12) and the full stain panel minus CD138. The images carry DOI [10.5281/zenodo.17381412](https://doi.org/10.5281/zenodo.17381412). [Canceridc](#) [GitHub](#)

Two caveats. IDC carries no clinical table for this collection, so age/sex above come from DICOM headers rather than a curated clinical file — and dates in de-identified DICOM are often shifted, so treat 2017-07-03 as an internal ordering reference rather than a true date. The matching clinical and genomic data (HIV status, sequencing, outcomes) live in the GDC under project CGCI-HTMC-DLBCL with the same submitter ID; I tried to pull that case record directly and the fetch kept returning a cached generic response, so I can't confirm those fields here. Much of the HTMC sequencing data is controlled-access through dbGaP anyway.

IDC skill/MCP configured Claude

This is a subject in **CGCI-HTMC-DLBCL** (IDC data release v24) — DICOM-converted whole slide images from the Cancer Genome Characterization Initiative's HIV+ Tumor Molecular Characterization Project for diffuse large B-cell lymphoma. It's pathology only; there's no radiology for this subject.

What's there

One study, dated 2017-07-03, described as "Histopathology": 12 slide microscopy (SM) series, 60 DICOM instances, about 36 GB. Each series is one digitized glass slide stored as a 5-level pyramid, scanned at 40× with 0.25 µm pixel spacing, brightfield. The images are large — roughly 113k-175k pixels wide by 80k-86k tall at full resolution.

The specimen is annotated as **lymph node**, with an admitting diagnosis code of *Diffuse large B-cell lymphoma, NOS*. Container IDs follow the pattern [HTMC-01-06-00611-01-<stain>](#).

The 12 stains — one H&E, ten IHC/ISH markers, plus an INT stain:

Slide	Stain
HE	hematoxylin + eosin
CD20, CD79a	B-cell markers
CD3	T-cell marker
CD10, BCL6, MUM1	Hans-algorithm cell-of-origin panel
BCL2, TP53, Ki-67	prognostic / proliferation
EBER	Epstein-Barr virus-encoded RNA probe
INT	iodonitrotetrazolium

That full panel makes this one of the more completely stained subjects in the collection — many of the 43 subjects lack the INT slide.

Access and terms

View in the browser (SLIM viewer, no download):

<https://viewer.imaging.datacommons.cancer.gov/slim/studies/2.25.11575476516094555195355863238822004089>

All 12 series are CC BY 4.0, so commercial use is fine with attribution. Cite: Clunie, D., Clifford, W., & Fedorov, A. (2026). *CGCI-HTMC-DLBCL: DICOM converted whole slide images from the Cancer Genome Characterization Initiative (CGCI) HIV+ Tumor Molecular Characterization Project (HTMC) - Diffuse Large B-Cell Lymphoma* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.17381412>

Two things worth flagging: IDC carries no clinical table for this collection, so HIV status, staging, treatment, and outcome aren't available here — the Zenodo record and the GDC entry for CGCI-HTMC-DLBCL are where supporting metadata lives. And at ~36 GB for the subject, you'd want to pull individual slides rather than the whole set if you're downloading. I can get you the S3/GCS URLs or an [idc download](#) manifest for whichever slides you want.

No IDC skill / MCP! ChatGPT

what can you tell me about subject HTMCP-01-06-00611 from IDC?

I can tell you that `HTMCP-01-06-00611` is a **GDC/CGCI subject identifier**, but I'm not finding an indexed public page that exposes the subject-level metadata directly.

What I can establish from IDC is:

- `HTMCP-01-06` belongs to the **HIV+ Tumor Molecular Characterization Project (HTMCP)**, part of NCI's Cancer Genome Characterization Initiative (CGCI).
- IDC currently has several HTMCP-derived collections, including:
 - **CGCI-HTMCP-CC** — cervical cancer, 211 subjects, whole-slide microscopy
 - **CGCI-HTMCP-LC** — lung cancer, 27 subjects, whole-slide microscopy
 - **CGCI-HTMCP-DLBCL** — diffuse large B-cell lymphoma, 43 subjects, whole-slide microscopy.
- The IDC images in these collections are **DICOM Slide Microscopy** conversions of the original Aperio SVS slides obtained from GDC.  +2

However, I **don't want to guess which HTMCP disease cohort** `00611` belongs to. The identifier itself isn't turning up in web indexing.

I can get much more specific

Since you're asking specifically about an **IDC subject**, the useful thing would be to query the IDC metadata for:

- collection
- patient/subject ID
- number of studies/series/instances
- modalities
- body part / anatomical site
- specimen and slide metadata
- stain / specimen description
- acquisition information
- links to the corresponding GDC case/sample
- potentially the actual DICOM series and image objects

If you want, I can ↪ **trace** `HTMCP-01-06-00611` **through IDC and GDC and give you the complete subject-level inventory**, including the exact slides/images available in IDC. 

No IDC skill / MCP!
Perplexity

what can you tell me about subject HTMCP-01-06-00611 from IDC?

»

Sources

10 >

Finished 2 steps >

HTMCP-01-06-00611 is an anonymized participant identifier from the *HIV+ Tumor Molecular Characterization Project–Cervical Cancer* cohort, commonly abbreviated **CGCI-HTMCP-CC**. In IDC, it refers to imaging associated with a cervical-cancer case in this research collection; it is not a clinical identifier that can be used to identify a person.

Study context

The HTMCP project was an NCI Cancer Genome Characterization Initiative effort studying cancers in people living with HIV, including cervical cancer. Its broad goal was to molecularly characterize HIV-associated tumors and compare them with analogous tumors from people without HIV. The cervical-cancer component was selected because cervical cancer is strongly HPV-associated and because suitable tissue was available for genomic analysis. [ncbi.nlm.nih](#)

The overarching dbGaP study ([phs000528](#)) reports **212 consented subjects** and a tumor-versus-matched-normal cohort design. Molecular profiling included paired tumor/germline whole-genome sequencing and tumor transcriptome sequencing in the wider program, alongside clinical and higher-level derived data released through NCI resources. [ncbi.nlm.nih](#)

What the IDC subject label means

The string likely encodes a study-specific de-identified case label:

Component	Likely role
HTMCP	HIV+ Tumor Molecular Characterization Project
01-06	Internal project/site or collection coding
00611	De-identified participant/case number

You should treat it as a **research-subject key**, useful for joining IDC imaging records to linked specimen, case, and potentially GDC genomic records where permitted—not as a code from which demographic, clinical stage, HIV status, treatment, or outcome can safely be inferred.

What may be available

For this subject, IDC may expose one or more image studies/series and associated DICOM metadata. Depending on the particular IDC release, useful fields commonly include:

No IDC skill / MCP!
Gemini

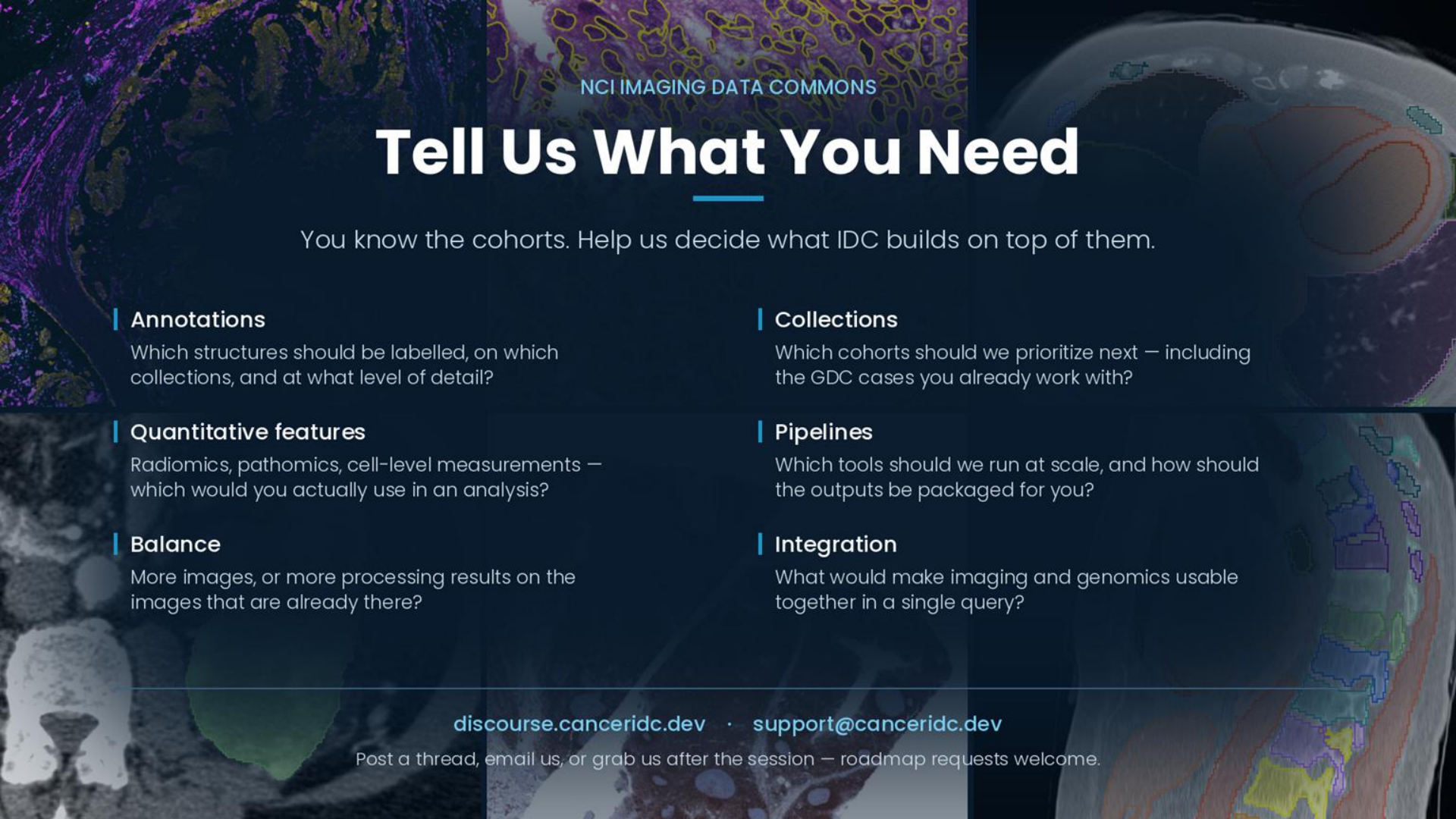
what can you tell me about subject HTMC-01-06-00611 from IDC?

Subject **HTMC-01-06-00611** is a specific patient case included in several major cancer research repositories, including the National Cancer Institute's Imaging Data Commons (IDC) and the Genomic Data Commons (GDC).

Based on the available identifiers and dataset structures, here is a breakdown of what this subject represents:

- **Project Origin:** The subject is part of the **CGCI-HTMC-DLBCL** project. This stands for the Cancer Genome Characterization Initiative (CGCI) – HIV+ Tumor Molecular Characterization Project (HTMC).
- **Disease Type:** The "DLBCL" designation means this patient was diagnosed with **Diffuse Large B-Cell Lymphoma**, an aggressive type of non-Hodgkin lymphoma.
- **Imaging Data (in IDC):** Within the NCI Imaging Data Commons, data for the CGCI-HTMC-DLBCL collection includes **DICOM-converted whole slide images (WSIs)**. This means pathology slides of this subject's tissue have been digitized, standardized, and made available for public research and machine learning analysis. [Zenodo](#)
- **Genomic Data (in cBioPortal/GDC):** This subject's molecular and clinical data (specifically under sample ID *HTMC-01-06-00611-01A*) is profiled in the cBioPortal for Cancer Genomics. It is included in the "**Mature B-Cell Neoplasms (Simon Fraser University, Blood 2023)**" study, which provides targeted sequencing and clinical annotations for DLBCL cases.

In summary, HTMC-01-06-00611 refers to an HIV-positive Diffuse Large B-Cell Lymphoma patient whose pathology imaging is hosted by the IDC, while their corresponding sequencing and clinical data can be queried through the GDC and cBioPortal to study the molecular drivers of their cancer.



NCI IMAGING DATA COMMONS

Tell Us What You Need

You know the cohorts. Help us decide what IDC builds on top of them.

Annotations

Which structures should be labelled, on which collections, and at what level of detail?

Quantitative features

Radiomics, pathomics, cell-level measurements — which would you actually use in an analysis?

Balance

More images, or more processing results on the images that are already there?

Collections

Which cohorts should we prioritize next — including the GDC cases you already work with?

Pipelines

Which tools should we run at scale, and how should the outputs be packaged for you?

Integration

What would make imaging and genomics usable together in a single query?

discourse.canceridc.dev · support@canceridc.dev

Post a thread, email us, or grab us after the session — roadmap requests welcome.