nature methods

Perspective

https://doi.org/10.1038/s41592-025-02643-0

Sharing data from the Human Tumor Atlas Network through standards, infrastructure and community engagement

Received: 11 July 2024

Accepted: 26 February 2025

Published online: 31 March 2025

Check for updates

Ino de Bruijn[1,7] ✉, Milen Nikolov[2,7], Clarisse Lau[3], Ashley Clayton[2], David L. Gibbs[3], Elvira Mitraka[2], Dar'ya Pozhidayeva[3], Alex Lash[4], Selcuk Onur Sumer[1], Jennifer Altreuter[4], Kristen Anton[5], Mialy DeFelice[2], Xiang Li[1], Aaron Lisman[1], William J. R. Longabaugh[3], Jeremy Muhlich[6], Sandro Santagata[6], Subhiksha Nandakumar[1], Peter K. Sorger[6], Christine Suver[2], Xengie Doan[2], Justin Guinney[2], Nikolaus Schultz[1], Adam J. Taylor[2], Vésteinn Thorsson[3], Ethan Cerami[4,8] ✉ & James A. Eddy[2,8]

Data from the first phase of the Human Tumor Atlas Network (HTAN) are now available, comprising 8,425 biospecimens from 2,042 research participants profiled with more than 20 molecular assays. The data were generated to study the evolution from precancerous to advanced disease. The HTAN Data Coordinating Center (DCC) has enabled their dissemination and effective reuse. We describe the diverse datasets, how to access them, data standards, underlying infrastructure and governance approaches, and our methods to sustain community engagement. HTAN data can be accessed through the HTAN Portal, explored in visualization tools—including CellxGene, Minerva and cBioPortal—and analyzed in the cloud through the NCI Cancer Research Data Commons. Infrastructure was developed to enable data ingestion and dissemination through the Synapse platform. The HTAN DCC's flexible and modular approach to sharing complex cancer research data offers valuable insights to other data-coordination efforts and researchers looking to leverage HTAN data.

The Human Tumor Atlas Network (HTAN) was launched by the National Cancer Institute (NCI) in September 2018, under the umbrella of the US Cancer Moonshot program. This initiative aims to accelerate cancer research and treatment, focusing on enabling scientific discovery, promoting collaboration and improving cancer data sharing[1]. HTAN is a step toward realizing these goals, with a mission to construct three-dimensional atlases of the dynamic cellular, morphological and molecular features of human cancers as they evolve from precancerous lesions to advanced diseases. As a consortium, HTAN aims to identify critical processes and events in the life cycle of human cancers, including the advancement of premalignant lesions to malignant tumors, the progression of malignant tumors to metastatic cancer, tumor response to therapeutic interventions, and the development of therapeutic resistance. In line with the broader goals of the Cancer Moonshot, HTAN is also committed to rapid, widespread data sharing with the wider scientific community (see Box 1 for a list of common terms).

[1]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Sage Bionetworks, Seattle, WA, USA. [3]Institute for Systems Biology, Seattle, WA, USA. [4]Dana-Farber Cancer Institute, Boston, MA, USA. [5]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [6]Harvard Medical School, Boston, MA, USA. [7]These authors contributed equally: Ino de Bruijn, Milen Nikolov. [8]These authors jointly supervised this work: Ethan Cerami, James A. Eddy. ✉e-mail: debruiji@mskcc.org; cerami@ds.dfci.harvard.edu

BOX 1

Glossary

ATAC-seq: assay for transposase-accessible chromatin sequencing, used to study chromatin accessibility.

Atlas: a collection of data focused on mapping the cellular and molecular characteristics of specific cancer types or stages.

BAM: binary alignment/map format, a binary file format used to store aligned sequence data.

cBioPortal: an open-source visualization tool for exploring multimodal cancer data.

CellxGene: an open-source visualization tool for visualizing single-cell RNA sequencing data.

CRDC: Cancer Research Data Commons, a network of cloud-based data repositories managed by the NCI.

DCC: Data Coordinating Center, manages the storage, standardization and sharing of HTAN data.

FASTQ: a file format for storing raw sequencing data, including nucleotide sequences and quality scores.

GDC: Genomic Data Commons, an NCI resource providing cancer researchers with access to genomic data.

H&E: hematoxylin and eosin staining, a common method for histological analysis of tissue samples.

HTAN: Human Tumor Atlas Network, a consortium focused on building three-dimensional atlases of the dynamic cellular, morphological and molecular features of human cancers as they evolve from precancerous lesions to advanced disease.

ISB-CGC: Institute for Systems Biology Cancer Gateway in the Cloud, a cloud-based platform for analyzing cancer genomics data.

Minerva: an open-source visualization tool for exploring multiplex imaging data.

OME-TIFF: a file format for storing high-resolution imaging data, commonly used in microscopy.

SB-CGC: Seven Bridges Cancer Genomics Cloud, a cloud-based resource for analyzing large cancer genomics datasets.

scRNA-seq: single-cell RNA sequencing, a method to measure gene expression at the individual cell level.

snRNA-seq: single-nucleus RNA sequencing, a technique to profile gene expression in the nucleus of cells.

Synapse: a data-sharing and collaboration platform developed by Sage Bionetworks.

TCGA: The Cancer Genome Atlas, a large-scale project that molecularly characterized more than 11,000 primary tumors across 33 cancer types.

TNP: Trans-Network Project, collaborative projects within HTAN focusing on specific research questions.

In the broader context of cancer research, HTAN draws on and extends The Cancer Genome Atlas (TCGA)[2], a landmark cancer genomics program that molecularly characterized more than 11,000 primary tumors and matched normal samples spanning 33 cancer types. TCGA generated comprehensive, multidimensional maps of the key genomic changes in major cancer types and subtypes, providing an invaluable resource for the cancer research community. HTAN is also part of a larger global effort to understand the human body at an unprecedented level of detail. Other initiatives, such as the Human Cell Atlas[3] and the Human BioMolecular Atlas Program consortium[4], are working to create comprehensive, high-resolution maps of all human cell types—healthy and diseased—as a basis for both understanding fundamental human biological processes and diagnosing, monitoring and treating disease. In a recent effort, the Curated Cancer Cell Atlas[5] published harmonized single-cell RNA sequencing (scRNA-seq) datasets to dissect intratumor heterogeneity. In comparison, HTAN is broader in scope, spanning many data types, and aims to provide well-annotated data to enhance similar resources and tools.

Earlier large cancer data-sharing efforts, such as TCGA, had their own complexities, but HTAN presents a new set of challenges. First, each HTAN atlas is unique and focused on exploring different hypotheses about cancer progression. As such, HTAN Centers (that is, grant recipients responsible for collecting and sharing data related to a particular tumor atlas research program) can use various experimental assays to support their study. They currently generate a highly diverse set of data types, including bulk sequencing, single-cell sequencing, multiplex imaging and spatial transcriptomics (Fig. 1a). Second, many of the experimental assays used within HTAN—particularly spatial profiling assays—are cutting-edge, necessitating that centers develop their own bioinformatics pipelines to perform analyses. Third, HTAN is focused on understanding temporal changes in cancer, and the HTAN data model must therefore be capable of capturing longitudinal clinical, phenotype and profiling data. Fourth, the multimodal nature of HTAN data requires multiple visualization and data-access resources, each of which must be tailored to individual data types or end-users.

To address the unique challenges of HTAN data, the network includes a dedicated Data Coordinating Center (DCC). The DCC is currently managed by personnel at four institutions: the Dana-Farber Cancer Institute, Sage Bionetworks, the Memorial Sloan Kettering Cancer Center and the Institute for Systems Biology. The DCC is responsible for developing HTAN data standards, managing HTAN data within a common cloud infrastructure, and sharing HTAN data with the scientific community (Fig. 1b). The DCC infrastructure includes centralized data ingestion, distributed data dissemination, user-friendly portals, and visualization tools. These activities are critical to ensuring that the wealth of data generated by HTAN is available for use by the broader scientific community.

The first phase of HTAN was completed in 2024. Here, we describe the diverse datasets generated and shared in this phase, the various methods for users to access HTAN data and metadata, the associated data standards, the technical infrastructure and governance approaches supporting the DCC, and ongoing community-engagement efforts.

Available data and data levels

HTAN data are now available for two pilot projects, ten atlases, and four trans-network projects (Supplementary Table 1). As of September 2024, this includes 2,088 research participants, 8,425 biospecimens and profiling data from more than 20 assays, covering a diverse range of techniques, including bulk, single-cell and spatial genomics; transcriptomics; epigenomics; hematoxylin and eosin staining; and multiplex imaging (Table 1). Clinical and biospecimen data have been collected and made available in tabular form. Assay data are organized into levels (Table 2), similar to previous TCGA efforts, with lower levels representing more raw data and higher levels corresponding to data processing by one or more bioinformatics pipelines. Each level for a particular data type adheres to a distinct, standard schema for file formats, metadata fields and values, as well as any additional data validation logic.

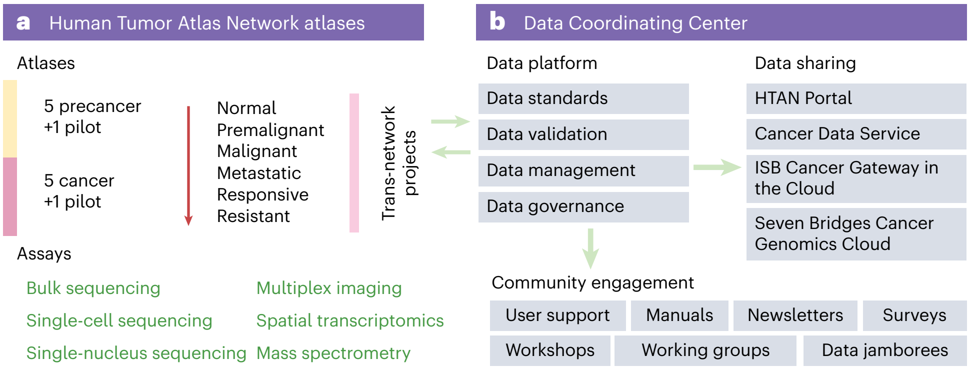


Fig. 1 | Overview of the HTAN Network and the HTAN Data Coordinating Center. a, HTAN atlases focus on specific transitions in cancer and generate a highly diverse set of data types. **b**, The HTAN DCC is responsible for developing data standards, managing data and sharing data with the scientific community.

Accessing data

HTAN data can be accessed through the HTAN Portal, as well as several services within the National Cancer Institute (NCI) Cancer Research Data Commons (CRDC)[6,7], such as the Institute for Systems Biology Cancer Gateway in the Cloud (ISB-CGC)[8], the Cancer Data Service (CDS) and the Seven Bridges Cancer Genomics Cloud (SB-CGC).

HTAN Portal

The primary mode of access is the dedicated HTAN Portal (https://humantumoratlas.org/) (Fig. 2a). The portal enables researchers to explore, access and download HTAN data through an intuitive user interface (UI). Users can specifically filter HTAN data through a number of criteria, including HTAN atlas, disease type, assay type or data level. User-friendly tools for advanced data querying and visualization are also available. The portal directs researchers to relevant routes of data access (Fig. 2b). Users can easily register for free and directly download open-access level 3 and 4 data from the Synapse data management platform (RRID SCR_006307). For controlled-access level 1 and 2 genomic and transcriptomic data, as well as level 2 imaging data, users are directed to corresponding data locations with the CDS. The portal also links to the HTAN Manual for more detailed information regarding the data model, tools and data repositories.

Visualizing and analyzing HTAN data

To enable seamless exploration of HTAN data, the HTAN Portal currently integrates multiple open source visualization and analysis tools (Fig. 2c). First, it integrates with Minerva, an open-source tool developed by Harvard Medical School for visualizing multiplex imaging data[9,10]. Two flavors of Minerva are currently supported: (1) Minerva Story, in which individual centers expertly annotate and describe specific data sets and delineate specific regions of interest; and (2) Auto-Minerva, which automatically generates Minerva images for all multiplex images and assigns reasonable channel defaults for viewing. Second, the portal integrates with cBioPortal for Cancer Genomics, an open-source tool for visualizing and analyzing cancer genomics data[11–13]. HTAN datasets including data from bulk sequencing and other methods, such as imaging or single-cell sequencing, are deposited into cBioPortal (https://cbioportal.org). Third, the portal integrates with CellxGene, an open-source tool developed by the Chan Zuckerberg Initiative for visualizing and analyzing single-cell data sets[14,15]. HTAN single-cell data are harmonized for deposition into CellxGene Discover (https://cellxgene.cziscience.com/), enabling exploration of HTAN data alongside other datasets in CellxGene Discover (see The HTAN Infrastructure Tooling).

Finally, HTAN data and metadata are available in ISB-CGC Google BigQuery. The platform features numerous BigQuery tables, including metadata tables, single-cell gene expression matrices and imaging channel data. We also provide several example notebooks to illustrate options for querying and analyzing HTAN data in ISB-CGC.

Controlled-access data

For controlled-access level 1 or 2 data, users must request access through the NIH database of Genotypes and Phenotypes (dbGaP, study accession phs002371). Once approved, users can access HTAN data in the cloud through SB-CGC. The HTAN Portal, ISB-CGC's Google BigQuery interface and CDS all provide the functionality to generate Data Repository Service (DRS)[16] manifest files for seamless access and analysis of HTAN data in SB-CGC. As of September 2024, there are 113 dbGaP-approved data-use plans for leveraging HTAN data for innovative applications. For instance, teams have integrated HTAN datasets with other genomic datasets to improve the detection of somatic and transcriptional alterations in cancers and to identify new biomarkers for early cancer diagnosis. Similarly, spatial transcriptomics and scRNA-seq data are being harnessed to identify cellular compositions and interactions within tumors, which could reveal new therapeutic targets and strategies. These data-reuse projects support the development of predictive models for disease progression and treatment response, ultimately contributing to personalized medicine and improved patient outcomes.

Data standards

HTAN has developed a unified data model that supports the management, standardization and exploration of clinical, biospecimen, molecular and imaging data across HTAN atlases. Clinical data encompass demographics, diagnosis, treatment, family history, environmental exposure and molecular test results. Biospecimen data capture information on storage conditions and provide end-to-end provenance from biopsy to acquisition. Assay metadata (those capturing experimental protocol and instrument context) include support for bulk and single-cell sequencing, multiplex imaging, and spatial transcriptomics. Complete details are available online via the HTAN Portal.

The HTAN data model was created and is maintained through a community-driven, peer-reviewed process. Members of a working group first assess established data standards and create a written request for comment (RFC) document soliciting community feedback. This document covers the data and all required and optional metadata elements, and usually undergoes several rounds of revision before all editors formally sign off on it. Through this process, the HTAN community has developed a consensus-driven data model that leverages multiple existing data standards and addresses community-driven use cases for data sharing and reuse. The HTAN data model specifically extends the clinical data model developed by the Genomic Data Commons (GDC)[17], the single-cell data model developed by the Human Cell Atlas[3] and the multiplex imaging model developed by the Minimum

Table 1 | Demographic, clinical and assay characteristics of HTAN participants

Characteristic	*n*=2,088		
Gender		**Tissue or organ of origin**	
Female	1,512 (72%)	Breast	945 (45%)
Male	460 (22%)	Lung	260 (12%)
Not Reported	116 (5.6%)	Pancreas	56 (2.7%)
		Colon	49 (2.3)
Race		Bone marrow	38 (1.8%)
White	1,450 (69%)	Sigmoid colon	35 (1.7%)
Black or African American	304 (15%)	Other	622 (30%)
Asian	41 (2.0%)	Not reported	205 (9.8%)
Other	10 (0.5%)		
Not reported	283 (14%)	**Assay**	
		Bulk DNA-seq	1,035 (50%)
Ethnicity		H&E	979 (47%
Not Hispanic or Latino	1,697 (81%)	Bulk RNA-seq	881 (42%)
Hispanic or Latino	41 (2.0%)	sc- or snRNA-seq	750 (36%)
Not reported	350 (17%)	Multiplexed tissue imaging	443 (21%)
		sc- or snATAC-seq	267 (12.6%)
Age at diagnosis (years, [25th–75th percentile])	51 [31, 63]	Spatial transcriptomics	232 (11.1%)
		Other	80 (3.8%)
Primary diagnosis			
Ductal carcinoma in situ	771 (37%)		
Adenocarcinoma	221 (11%)		
Ductal carcinoma	102 (4.9%)		
Malignant melanoma	66 (3.2%)		
Carcinoma	60 (2.9%)		
Neuroblastoma	56 (2.7%)		
Other	396 (19%)		
Not reported	325 (16%)		

DNA-seq, DNA sequencing; H&E, hematoxylin and eosin staining; RNA-seq, RNA sequencing; sc- or snRNA-seq, single-cell or single-nucleus RNA-seq; sc- or snATAC-seq, single-cell or single-nucleus assay for transposase-accessible chromatin sequencing. More details can be found on the HTAN Portal.

Table 2 | Levels of HTAN data

Level	Single-cell RNA-seq	Multiplex imaging	Spatial transcriptomics
1	Unaligned sequencing reads, usually in the FASTQ file format.	Raw imaging tiles that require preprocessing, such as stitching, registration or background subtraction. Typically TIFF or a proprietary format	Unaligned sequencing reads, usually in the FASTQ file format.
2	Aligned sequencing reads, usually in the BAM file format.	Multichannel image. Usually in the OME-TIFF file format, accompanied by a CSV file containing channel metadata.	Aligned sequencing reads, usually in the BAM file format.
3	Gene expression matrix. For example, a matrix of all cells by all genes, with expression counts. Multiple file formats are supported, including CSV, MTX and h5ad.	Segmentation masks denoting nuclei, cytoplasm, whole cells or regions of interest. Multiple file formats are supported although TIFF and OME-TIFF are recommended.	Gene expression matrix. For example, a matrix of all cells by all genes, with expression counts. Multiple file formats are supported, including CSV, MTX and h5ad.
4	Feature matrix. For example, a matrix of cluster assignments or imputed cell types across all sequenced cells. Multiple file formats are supported, including CSV and h5ad.	Feature matrix. For example, a matrix of mean intensity values per cell and channel Multiple file formats are supported, including CSV and h5ad.	Feature matrix. For example, a matrix of cluster assignments or imputed cell types across all sequenced cells. Multiple file formats are supported, including CSV and h5ad.

Lower levels indicate raw data and higher levels indicate data analyzed by one or more bioinformatics or image-processing pipelines. Three primary categories of data are highlighted.

Information about Highly Multiplexed Tissue Imaging (MITI) consortium[18]. The data model is continuously evolving and refined on the basis of feedback from the reuse of HTAN data, as well as the introduction of new assays by data submitters.

The HTAN data model is formally represented as an open-access, extensible JSON-LD schema document (https://json-ld.org), enabling version control, linking of individual data elements to existing NCI data standards, and the creation of automated validation tools. The JSON-LD schema uses the Schema.org specification, which, in the case of HTAN, allowed a data model to be built by reusing existing biomedical ontologies when feasible, while adding new HTAN-specific extensions as needed. This promotes interoperability by reusing data elements for experimental variables shared across consortia. It also enhances downstream data discovery through services such as Google Datasets Search[19].

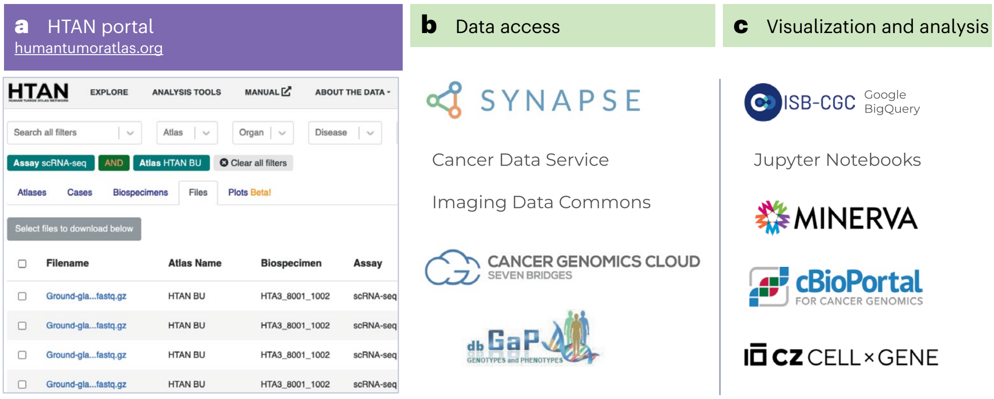
The model comprises more than 1,000 attributes across more than 30 modalities, analysis and data-processing types. A set of 113 HTAN common data elements is available in the NCI Cancer Data Standards Registry and Repository (caDSR)[20], ensuring that these data elements are available to the scientific community through the caDSR portal, application programming interface and tools. These data elements may be collectively browsed and retrieved under the HTAN classification.

Infrastructure

A broad range of tools, data standards and platforms have been leveraged, enhanced or developed to support the overall HTAN DCC data infrastructure. This includes tooling to support data and metadata ingestion, data storage, access controls, quality assurance, data sharing, image processing, visualization and analysis (Supplementary Table 2). All data standards and most tools are available at GitHub (https://github.com/ncihtan) and the HTAN Portal's Tools Page, and are freely available to other consortia that wish to build on the work of HTAN.

Governance and policy

Responsible data sharing requires clear governance to ensure that data contributors, curators and users can share and use data effectively. The DCC collaborates with the HTAN consortium to create data-sharing agreements and policies on the basis of the NCI Cancer Moonshot Public Access and Data Sharing Policy[21]. These policies outline the conditions under which HTAN data are made public and how institutions unaffiliated with HTAN can contribute data using HTAN services. The Synapse platform supports and enforces these policies by managing team-level



□	Filename	Atlas Name	Biospecimen	Assay
□	Ground-gla...fastq.gz	HTAN BU	HTA3_8001_1002	scRNA-seq
□	Ground-gla...fastq.gz	HTAN BU	HTA3_8001_1002	scRNA-seq
□	Ground-gla...fastq.gz	HTAN BU	HTA3_8001_1002	scRNA-seq
□	Ground-gla...fastq.gz	HTAN BU	HTA3_8001_1002	scRNA-seq

Fig. 2 | The HTAN Portal. a, The query interface for finding data and tools. **b**, Data access recipes for lower level (1–2) and higher-level (3–4) data. **c**, Visualization and analysis tools for exploring HTAN data.

access controls, ensuring that HTAN centers, data users and DCC staff have appropriate data access. Governance experts from the DCC played a key role in HTAN's policy working group, aligning the HTAN research community and ensuring policy consensus—a prerequisite for HTAN's data-sharing success.

To protect the privacy of research participants, the HTAN data-sharing policy requires that HTAN Centers deidentify data before submitting them to the DCC through Synapse. The DCC conducts further modality-specific checks to ensure privacy in data derivatives. This includes executing policies to detect and remove date information from imaging data that could be used to reconstruct sensitive data, such as birth dates.

Additional policies cover publications, research protocols and computational tools, all accessible on the HTAN Portal, as resources for the HTAN community and other DCC programs.

Community engagement

As with any large-scale scientific consortium, it is essential to ensure both transparent communication and coordination among principal investigators, data contributors, and method and tool developers, as well as other key stakeholders, and engagement with the wider scientific community. In the consortium, the DCC works to engage all HTAN members at multiple levels. This includes biannual face-to-face meetings, junior-investigator workshops, data workshops and working groups devoted to policy implementation and scientific collaboration. As noted previously, working groups also drive the RFC process for developing and evolving the HTAN data model. There were 136 non-DCC HTAN representatives who contributed across 18 data standard RFCs, providing 871 comments.

DCC staff provide support for specific HTAN Centers as liaisons, covering technical areas such as imaging data or clinical metadata. These data liaisons serve as designated points of contact and facilitate communication between the contributing HTAN Centers and the DCC. Private channels on the messaging tool Slack and a help desk ensure that data contributors can engage with the DCC about inquiries and to track bugs or submission issues.

The DCC actively engages the broader scientific community through timely data releases, outreach to other scientific consortia, and public workshops such as data jamborees and scientific conferences. The jamborees have been particularly helpful in providing feedback on data accessibility. For example, jamboree participants have identified issues in finding specific samples from publications or identifying HTAN data in the CGC, which we then improved. We also actively maintain an HTAN manual (https://docs.humantumoratlas.org), our primary external documentation, designed to explain the consortium to new users. The manual describes available HTAN data, HTAN data standards and all modes of data access. A publicly accessible HTAN Help Desk is open to external researchers with data-specific questions. Finally, we ensure that HTAN data are available through multiple channels across the NCI cancer data ecosystem through the CRDC[6,7].

The HTAN infrastructure tooling

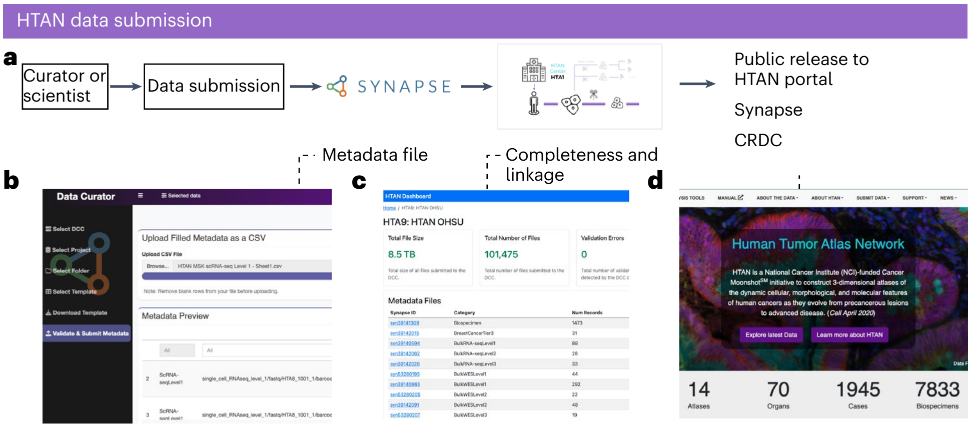
The HTAN data-submission process

The DCC has developed a standardized data-submission process (Fig. 3a). The process begins when a data curator or scientist at an HTAN Center uploads their data to cloud buckets connected to Synapse. Once the data are uploaded, the submitter needs to provide metadata for each file, including information about its processing and the research participant and biospecimen that it applies to. These metadata are critical for data access and reuse. Metadata are submitted through the Data Curator App (DCA) (Fig. 3b), which creates a metadata template on the basis of the data model, validates the provided metadata against the data model, and uploads it to Synapse. Centers also have the option of submitting a filled metadata template describing individual publications and all data associated with a publication.

After submitting metadata, a second set of validation checks is automatically performed. These checks examine the HTAN Center's dataset as a whole, verify that all assay data can be linked to parent biospecimens and research participants, and assess data for overall completeness. The results of these checks are made available through the HTAN Dashboard (Fig. 3c).

After a new data submission, HTAN DCC members review the HTAN Dashboard and relay validation issues to the data submitters at the respective HTAN Center. This feedback cycle continues until all validation errors are resolved. Once both the DCC and the center sign off, all files intended for release are queued. An HTAN Portal preview instance is generated with all data for the next release. After a final manual check, all release data are deployed to the public HTAN Portal. Higher-level processed data are made publicly available on Synapse. Lower-level access-controlled data are submitted to the CRDC[6,7], where they are made available in subsequent CRDC releases. Data are also submitted in a parallel process to other platforms, including CellxGene[14], cBioPortal[11–13] and ISB-CGC[8], each with its own release cycle. A future goal is to automate this broader dissemination process.

Setting deadlines for major data releases helps to incentivize centers to submit data in a timely manner. Major releases are completed twice per year, with minor releases on an as-needed basis. A complete log of data releases is maintained on the HTAN Portal. Although HTAN aims to release data upon generation, in practice, we have found that most centers submit data closer to manuscript submission as incentivized by publishers' data access requirements and the desire to ensure high data quality before release.



Upload Filled Metadata as a CSV

Upload CSV File

HTAN MSK scRNA-seq Level 1 - Sheet1.csv

Note: Remove blank rows from your file before uploading.

Metadata Preview

2	ScRNA-seqLevel1	single_cell_RNAseq_level_1/fastq/HTA8_1001_1/barcod
3	ScRNA-seqLevel1	single_cell_RNAseq_level_1/fastq/HTA8_1001_1/barcod

HTA9: HTAN OHSU

Total File Size	Total Number of Files	Validation Errors
8.5 TB	101,475	0
Total size of all files submitted to the DCC.	Total number of files submitted to the DCC.	Total number of validat detected by the DCC c

Metadata Files

Synapse ID	Category	Num Records
syn39141309	Biospecimen	1473
syn39142015	BreastCancerTier3	31
syn39140594	BulkRNA-seqLevel1	88
syn39142062	BulkRNA-seqLevel2	28
syn39142026	BulkRNA-seqLevel3	33
syn53280193	BulkWESLevel1	44
syn39140863	BulkWESLevel1	292
syn53280205	BulkWESLevel2	22
syn39142091	BulkWESLevel2	48
syn53280207	BulkWESLevel3	19

Human Tumor Atlas Network

HTAN is a National Cancer Institute (NCI)-funded Cancer Moonshot^{SM} initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease. (*Cell April 2020*)

14	70	1945	7833
Atlases	Organs	Cases	Biospecimens

Fig. 3 | HTAN Data submission and release process. a, An HTAN data curator or scientist uploads data to AWS, Google Cloud or Synapse, provides metadata about each file, and confirms metadata validation. The DCC performs additional quality-control checks and releases data to the public. **b**, The DCA performs metadata validation. **c**, The HTAN Dashboard performs additional quality-control data checks and checks for overall data completeness. **d**, The DCC releases the data to the public.

Synapse

Sage Bionetworks uses its data-management platform, Synapse (RRID SCR_006307), as the central repository for the HTAN DCC. Each HTAN Center has a dedicated Synapse project, providing a secure environment for uploading, organizing and annotating data and metadata before public release. Synapse streamlines this process through multiple features, including wikis, entity annotations, tabular annotation views for file exploration, and finely tuned access control settings, creating a user- and machine-friendly data-management ecosystem.

Project access on Synapse is regulated through team membership, with adjustable permission levels to ensure appropriate access for both data contributors and DCC staff. Moreover, HTAN's Synapse projects integrate with external storage solutions, such as AWS S3 and Google Cloud Storage, allowing Centers to choose their preferred storage provider, which can minimize egress costs. This is particularly advantageous for contributors who already have data stored with these providers. The platform supports the synchronization of directly added storage objects into Synapse using serverless architectures, such as AWS Lambda and Google Cloud Functions. This integration facilitates efficient data uploads through cloud provider clients while preserving the user-friendly experience of Synapse's web UI, command line interface and language-specific clients in Python and R. For HTAN, the only requirement around folder structure for each center is that all submissions are grouped into top-level folders categorized by data type, such as scRNA-seq FASTQ files, imaging OME-TIFFs or demographic information. File naming is minimally restrictive because essential information is captured in the metadata rather than the file names themselves.

Data curator app

The DCA (Fig. 3b), hosted on AWS Fargate, enables data submitters to associate metadata with their assay data files through a wizard-style interface in the browser. The application backend leverages a Python tool, Schematic, to validate the metadata files against the HTAN data standards and submit data to Synapse. Both DCA and Schematic were developed to support multiple data-coordination projects at Sage Bionetworks. The separation of UI (DCA) and programmatic schema validation logic (Schematic) simplifies the reuse of these tools across different projects.

In the metadata submission wizard, data contributors select a template (for example, metadata for clinical demographics or level 1 scRNA-seq). A Google Sheets link is generated, allowing users to directly fill out the metadata template online using Google Sheets' functionalities. The Google Sheets template includes checks for the correctness of particular columns. If preferred, the sheet can also be exported as a delimited text file or Excel spreadsheet. Should a specific template be unavailable, a minimal metadata template is used, with the provision to contact a DCC liaison for further guidance. After completing the template, users submit it, and the DCA then uses Schematic to do an additional check for schema correctness and submits it to Synapse. DCA also allows for existing metadata to be updated, accommodating corrections, compliance adjustments or additions for new files.

HTAN Dashboard

The HTAN Dashboard (Fig. 3c) is a web application developed to help data submitters across the HTAN Centers and the DCC to track submitted data and associated metadata. For each HTAN Center, the dashboard performs various checks, including tracing and validating all links from files to samples to research participants and ensuring that HTAN ID numbers adhere to specified guidelines. It also calculates a metadata completeness score to assess how complete the provided metadata are in terms of supplied values compared with empty fields. The dashboard provides summary statistics, including file counts and sizes per atlas and the number of remaining data submission errors. The HTAN Dashboard is written in Python and leverages the Synapse client to programmatically retrieve each center's metadata and file counts.

Image visualization on the HTAN Portal

HTAN Centers generate imaging data using a broad array of multiplex imaging assays. To enable initial visualization and exploration of these data directly on the HTAN Portal, we deployed narrative guides using Minerva, a lightweight tool suite for interactive viewing and fast sharing of large image data[9]. Although extensively curated and interactive guides with manual channel thresholds, waypoints and regions of interest can be generated, we implemented an automatic channel thresholding and grouping approach to generate good first defaults, enabling the rapid generation of prerendered Minerva stories, which can be enhanced with interactive channel selection and embedded metadata. To facilitate recognition and recall of images and tissue features from multiplexed tissue images, we developed Miniature, a new approach for creating informative and visually appealing thumbnails from multiplexed tissue images.

HTAN data in CZ CellxGene discover

Single-cell sequencing data are submitted to CZ CellxGene Discover. The platform enables users to find, explore, visualize and analyze published datasets. To ensure integration with other single-cell datasets, HTAN data are harmonized to adhere to the CellxGene schema and data format requirements. The HTAN data-ingestion workflow captures much of the same information, including raw counts, normalized counts, demographics (for example age, sex and ethnicity), assay type, tissue site, disease type and embeddings (for example uniform manifold approximation and projection and *t*-distributed stochastic neighbor embedding). A key additional requirement is the annotation of cell types using terms from the Cell Ontology initiative (CL, https://obofoundry.org/ontology/cl), which currently is performed by manual mapping of annotations (cell phenotypes) provided by data contributors to the closest CL terms. For example, there was no term for lymphomyeloid primed progenitor-like blasts[22]; instead, hematopoietic multipotent progenitor cell (CL_0000837) was selected. Precancer- and cancer-cell mapping posed a challenge because CL is largely based on classifications of normal cells. Cancer cells are annotated according to the presumed healthy cell type from which they originated. For cases in which no appropriate cell type terms are available, the most relevant parent ontology is used to describe the cell type. The CL version is 2024-04-05, based on CellXGene's v1.0.5 schema requirements. We curated 17 HTAN datasets for CellxGene. In general, we found that data submitters are willing to do this additional work to facilitate the reuse of their data. We plan to provide cell-type annotations for all HTAN single-cell data submissions in the future, manually or through automated pipelines, and reannotate them as CL's coverage and quality improve.

Integration with the Cancer Research Data Commons

HTAN data ingress and standardization processes are integrated with the Cancer Research Data Commons (CRDC) ecosystem, with multiple services supporting HTAN data download, queries and processing. Specifically, CDS provides access to HTAN controlled-access sequence and imaging files; SB-CGC provides mechanisms to run a variety of processing workflows on HTAN data at CDS; and ISB-CGC contains HTAN tabular metadata and assay data for flexible queries.

HTAN imaging data are available through CDS in original contributed formats, including OME-TIFF and SVS files. Preserving contributor-provided formats facilitates both reproducibility of published studies and interoperability with common processing and visualization tools, including processing suites, such as MCMICRO[23], and analysis tools, such as Napari[24] and QuPath[25]. A subset of HTAN imaging data has been deposited in the NCI's Imaging Data Commons[26], where data have been converted to DICOM[27] to provide interoperability with other medical imaging datasets and tools.

The NCI's cloud resources allow processing of HTAN data on the cloud. For example, SB-CGC[28] facilitates selection and processing of HTAN scRNA-seq read-level files, image data files and read-level spatial transcriptomic data. Within ISB-CGC[8], HTAN data are available as Google BigQuery tables, allowing flexible SQL query access. More than 850 assay files are queryable through Google BigQuery, encapsulating data from imaging level 4 and scRNA-seq level 4 assays, collectively spanning more than 200 million cells across spatial and single-cell datasets. Computational notebooks are provided to illustrate cloud-based querying and processing of HTAN data.

Discussion

As of September 2024, HTAN is planned to continue for at least another 5 years. We developed a flexible and modular open-source infrastructure to ingest and disseminate data, enabling coevolution with emerging assay technologies and expanding data capture in the clinic. We believe the approaches used by HTAN will be useful for data coordinating centers of other consortia, and we have already seen aspects of it reused in other more recently formed consortia, including the Gray Foundation BRCA Pre-Cancer Atlas[29] and the Break Through Cancer Foundation, as well as across other data repositories such as the CRDC. Although the HTAN data resource is an aggregation of unique hypothesis-driven studies with context-specific experimental design, there is potential for pan-cancer analyses owing to overlap in employed assays both in and outside of HTAN. For instance, single-cell data can be used to identify gene expression patterns across tumor types or expression of a particular gene in the HTAN data could be compared with that in other CellxGene datasets from healthy tissues. Other examples from the data jamborees include improving image-segmentation algorithms, identifying markers of tumor progression in transcriptomics data, and comparing cell-type identification across assay methods. Improvements in data harmonization tooling will benefit these use cases. Because there is now a wealth of HTAN data available, we plan to continue to engage with the community through tutorials, webinars and data jamborees, and streamline the reuse of HTAN data on the basis of user feedback. More data will be collected and integrated to further improve the utility of HTAN data. The new data will include improvements to sample collection, for example incorporating more tumor types and a more diverse cohort of participants, as well as more precise and seamless recording of which protocols and data-processing methods were used by each center. Similar assays, sample-collection methods and data-processing techniques across tumor types could further benefit pan-cancer analyses. Our infrastructure roadmap includes improvements to data ingestion (for example, additional data integrity checking), data harmonization, the data-release process (increased automation and improved data tracking), and dissemination through the HTAN Portal (enhanced publication pages) and the broader cancer data ecosystem, including streamlined releases to CRDC, CellxGene, cBioPortal and other repositories.

Data availability

All data are available via the HTAN Portal: https://humantumoratlas.org.

Code availability

All data standards and tools are available via GitHub (https://github.com/ncihtan) and the tools page on the HTAN Portal (https://humantumoratlas.org). A detailed list of tooling and corresponding repositories is provided in (Supplementary Table 2).

References

1. Sharpless, N. E. & Singer, D. S. Progress and potential: the Cancer Moonshot. *Cancer Cell* **39**, 889–894 (2021).
2. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
3. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
4. Jain, S. et al. Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). *Nat. Cell Biol.* **25**, 1089–1100 (2023).
5. Gavish, A. et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
6. Hinkson, I. V. et al. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell Dev. Biol.* **5**, 83 (2017).
7. Wang, Z. et al. NCI Cancer Research Data Commons: resources to share key cancer data. *Cancer Res.* **84**, 1388–1395 (2024).
8. Reynolds, S. M. et al. The ISB Cancer Genomics Cloud: a flexible cloud-based platform for cancer genomics research. *Cancer Res.* **77**, e7–e10 (2017).
9. Hoffer, J. et al. Minerva: a light-weight, narrative image browser for multiplexed tissue images. *J. Open Source Softw.* **5**, 2579 (2020).
10. Rashid, R et al. Narrative online guides for the interpretation of digital-pathology images and tissue-atlas data. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-021-00789-8 (2021).
11. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
12. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

13. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
14. Megill, C. et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.05.438318 (2021).
15. CZI Single-Cell Biology et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.30.563174 (2023).
16. Thorogood, A. et al. International federation of genomic medicine databases using GA4GH standards. *Cell Genom.* **1**, 100032 (2021).
17. Heath, A. P. et al. The NCI Genomic Data Commons. *Nat. Genet.* **53**, 257–262 (2021).
18. Schapiro, D. et al. MITI minimum information guidelines for highly multiplexed tissue images. *Nat. Methods* **19**, 262–267 (2022).
19. Benjelloun, O., Chen, S. & Noy, N. Google Dataset search by the numbers. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2006.06894 (2020).
20. Warzel, D. B. et al. Common data element (CDE) management and deployment in clinical trials. *AMIA Annu. Symp. Proc.* **2003**, 1048 (2003).
21. NCI. Cancer Moonshot^{SM} public access and data sharing policy. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/funding/public-access-policy (National Cancer Institute, 2021).
22. Chen, C. et al. Single-cell multiomics reveals increased plasticity, resistant populations, and stem-cell-like blasts in KMT2A-rearranged leukemia. *Blood* **139**, 2198–2211 (2022).
23. Schapiro, D. et al. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. *Nat. Methods* **19**, 311–315 (2022).
24. Napari Contributors. napari: a multi-dimensional image viewer for python. *Zenodo* https://doi.org/10.5281/zenodo.3555620 (2019).
25. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci Rep.* **7**, 16878 (2017).
26. Fedorov, A. et al. NCI imaging data commons. *Cancer Res.* **81**, 4188–4193 (2021).
27. National Electrical Manufacturers Association. *Digital Imaging and Communications in Medicine (DICOM) Standard* Report No. NEMA PS3 / ISO 12052 (DICOM, 2025).
28. Lau, J. W. et al. The Cancer Genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer Res.* **77**, e3–e6 (2017).
29. Kader, T. et al. Multimodal spatial profiling reveals immune suppression and microenvironment remodeling in fallopian tube precursors to high-grade serous ovarian carcinoma. *Cancer Discov.* https://doi.org/10.1158/2159-8290 (2024).

Acknowledgements

The HTAN DCC is supported by NCI Grant U24CA233243 (principal investigators E.C., N.S., V.T., J.A.E., A.J.T.). Support for this work was provided to Memorial Sloan Kettering Cancer Center by a core grant from the National Cancer Institute (P30 CA008748). We thank all research participants for their contributions and the HTAN Centers for collecting and providing the data. We acknowledge the contributions of DCC staff past and present, including A. Abeshouse, E. Kozlowski, L. Williams, J. Hwee, D. Gutman, S. Reynolds, P. Kumari, A. Gopalan, B. Zalmanek, T. Adams, J. Vera, T. Yu, A. Heiser, B. Macdonald and Y. Chae.

Author contributions

These authors contributed equally: I.d.B., M.N. I.d.B., M.N., C.L., D.L.G., E.M., D.P., W.J.R.L., N.S., A.J.T., E.C., V. T. developed data standards; I.d.B., M.N., E.M., M.D., A.J.T, S.O.S., C.L, X.L., X.D., A.L., S.N., W.J.R.L., V.T., E.C. developed DCC infrastructure; I.D.B, M.N., C.L., D.L.G., D.P., A.L., J.A., K.A., X.L., S.N., N.S., A.J.T., V.T., E.C. served as data liaisons: J.M. S.S. P.K.S., A.J.T.: developed tooling for imaging data; C.S., W.J.R.L. developed data sharing and governance policies; A.C., A.L.: coordinated projects, meetings and communication; I.D.B, M.N., C.L., A.C., D.L.G., D.P., A.L., J.A., K.A., S.N., C.S., N.S., A.J.T., V.T., E.C., J.A.E. wrote the manuscript; I.D.B, M.N., C.L., A.C., D.L.G., E.M., D.P., A.L., S.O.S., J.A., K.A., M.D., X.L., A.L., W.J.R.L, J.M., S.S., S.N., P.K.S., C.S., X.D., J.G., N.S., A.J.T., V.T., E.C., J.A.E. edited and reviewed the manuscript; N.S., A.J.T., V.T., J.G., E.C., J.A.E supervised the HTAN DCC.

Competing interests

P.K.S. is a cofounder and member of the BOD of Glencoe Software, member of the BOD for Applied BioMath and a member of the SAB for RareCyte, NanoString, Reverb Therapeutics and Montai Health; he holds equity in Glencoe, Applied BioMath and RareCyte. S.S. is a consultant for RareCyte. The other authors declare no competing interests

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02643-0.

Correspondence should be addressed to Ino de Bruijn or Ethan Cerami.

Peer review information : *Nature Methods* thanks Stefano Mangiola, Casey Greene and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lei Tang, in collaboration with the *Nature Methods* team.