Constructing and Using Cell Type Populations of the Human Reference Atlas

Andreas Bueckle[1]*, Bruce W. Herr II[1]*, Lu Chen[2], Daniel Bolin[1], Danial Qaurooni[1], Michael Ginda[1], Yashvardhan Jain[1], Aleix Puig-Barbe[3], Kristin Ardlie[4], Fusheng Wang[2], Katy Börner[1]*

[1] Department of Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, 47408, USA
[2] Department of Computer Science and Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, 11794, USA
[3] European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[4] Broad Institute, Cambridge, MA, 02142, USA
* Corresponding authors
abueckle@iu.edu
bherr@iu.edu
katy@iu.edu

Abstract

The human body contains ~27–36 trillion cells of up to 10,000 cell types (CTs) within a volume of ~62–120 liters (males) and 52–89 liters (females). The Human Reference Atlas (HRA) v2.3 provides a quantitative 3D framework of CTs across 73 reference organs and 1,283 3D anatomical structures (ASs). The HRA Cell Type Population (HRApop) effort quantifies CTs per AS using high-quality single-cell (sc) data processed through scalable, reproducible workflows and cell type annotation (CTann) tools. HRApop v1.0 includes reference CT populations for 73 ASs (112 when sex-specific) using 662 datasets spatially registered to 230 locations across 17 organs (31 when sex-specific). For 558 sc-transcriptomics datasets (11,042,750 cells), CTs and biomarker expression were computed using Azimuth, CellTypist, and popV. To test generalizability, 104 sc-proteomics datasets (16,576,863 cells) were integrated. In total, HRApop includes 27,619,613 cells. HRApop can be used to predict (1) CT populations for 3D volumes in the human body and (2) the spatial origin of a tissue block, given a CT population. Data and code are at cns-iu.github.io/hra-cell-type-populations-supporting-information.

Background & Summary

The Need. The volume[1] of the adult human body is estimated to range from 62-120 liters (0.062-0.120 m³) in males with 36 or 37[2] trillion cells to 52-89 liters (0.052-0.089 m³) in females with 27 trillion cells[3]. There is no consensus on the number of CTs within the human body. Estimates range from 400 major CTs[3–6] to 3,358[7] total CTs, and depend on the criteria used to determine what constitutes a CT (see the **Methods** section for a more detailed discussion). Efforts like the Human BioMolecular Atlas Program (HuBMAP)[8,9], Human Cell Atlas (HCA)[10–12], and many of the 20+ other atlas efforts contributing to the HRA endeavor aim to provide clarity based on high-quality experimental data collection and analysis.

Most atlas projects aim to capture the number and type of cells per AS together with biomarker expression values—based on expert knowledge or experimental data. For example, the Blue Brain Cell Atlas[13] (bbp.epfl.ch/nexus/cell-atlas) features 3D data for the mouse brain with CT populations programmatically placed inside 737 brain regions defined in the Allen Mouse Brain Atlas[14]; different ASs; regions and their CT populations can be toggled on and off, and the color hue can be set to encode cell regions, types, or density;

cell counts for neurons and glia (with confidence values) are also displayed. The Genotype-Tissue Expression (GTEx) sc-portal[15] (gtexportal.org/home/singleCellOverviewPage) features CT populations from 25 tissue blocks in eight organs from 16 donors, and the Chan Zuckerberg Initiative (CZI) CELLxGENE (CxG) Portal[16] (cellxgene.cziscience.com/datasets) features CT populations for 325 human, organ-level ASs but without visual representations of these structures. At the functional tissue unit (FTU) level[17], Anatomograms[18,19] developed by the Gene Expression Team at the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) in collaboration with the Wellcome Sanger Institute; the Kidney Precision Medicine Project (KPMP)[20,21] Explorer (atlas.kpmp.org/explorer); and the HRA FTU Explorer (apps.humanatlas.io/ftu-explorer) feature 2D medical illustrations with associated CT populations.

Given that life unfolds in 3D, there is a strong interest to capture CT populations for ASs in the human body in 3D. The HRA v2.3 features the 3D shape, size, location, and rotation of 1,283 3D ASs, which can be explored at humanatlas.io/3d-reference-library; each 3D AS belongs to one of 73 organs, also called 3D reference objects (see **Box 1** in a related publication[22]). As new organ expert teams join the HRA effort, new 3D structures for male and female are added. Data from sc-portals such as GTEx, CxG, KPMP (atlas.kpmp.org), the HuBMAP Data Portal (portal.hubmapconsortium.org), or the Cellular Senescence Network (SenNet) Data Portal[23] (data.sennetconsortium.org/search) can be spatially registered into the 3D reference objects using the web-deployed HRA Registration User Interface (RUI, apps.humanatlas.io/rui)[24,25] and guided by standard operating procedures (SOPs)[26–28].

Ultimately, HRApop can serve as a healthy reference to elucidate mechanisms underlying cellular interactions and, as a result, point to changes in the body that signify disease progression, which may facilitate advancements in basic discovery and lead to new therapeutic strategies.

The Challenge. Computation of CT populations for the many different ASs in the human body requires both a 3D registration and reproducible dissociation protocols that allow isolating single cells (or single nuclei) from tissue in support of sc-transcriptomics analyses[29–31] or high quality cell segmentation for sc-proteomics spatial data[32,33]. In both cases, CTann tools are needed to assign a CT to each cell. Most sc-segmentation and annotation tools are organ specific. However, human organs are large (an average human kidney is about 10-12 cm long, 5-7 cm wide, and 3 cm thick and is estimated to contain about 110 billion cells[3]). Organs have many different internal ASs with vastly different CT populations that serve diverse physiological functions. The HRA[34,35] used the Visible Human Project data[35–38] and the expertise of medical illustrators to segment 1,283 3D ASs in 37 vital organs; given tissue data that is spatially registered into these 3D AS using the RUI[24] plus sc-dissociation/segmentations and annotations[32,39–47], the number of cells per CT can be computed for specific ASs, i.e., 15 AS within the male lung. To compute HRApop at AS (not organ level), 1,000s of datasets from four different portals are downloaded, donor metadata is harmonized, and multiple CTann tools are run for each sc-transcriptomics dataset. This paper explains exactly how dataset, extraction site, and AS-specific CT populations and biomarker expression values are computed at scale and how they are used to predict CT populations for 617 tissue blocks and spatial origin for 5,453 datasets in support of quantifying cross-AS and cross-individual variation of cellular phenotypes; how open HRApop v1.0 data is published via the HRA; and what known limitations and next steps exist for the work presented.

Fig. 1 shows the 112 ASs of the male (left) and female (right) reference body for which spatially registered sc-transcriptomics data exist. For each AS, the organ name, number of datasets, and AS name plus a bar graph with the percentage of major CTs, so-called AS Cell Type Populations (ASpop, see **Box 1**), are shown. Only sc-transcriptomics data is shown and three CTann tools were used to annotate cells: Azimuth[48], CellTypist[49,50], and/or popV[51]; with a preference for Azimuth annotation over CellTypist over popV[51], see **Methods** section for details. Crosswalking to Cell Ontology (CL, www.ebi.ac.uk/ols4/ontologies/cl)[52–54] and aggregation to higher-level CTs is also described in detail in the **Methods section**. For the left and right mammary gland, two CTs

that cannot be mapped to a high-level CT make up ~70% of the CTs across ASs. As a result, the stacked bar graphs for these are mostly grey (luminal epithelial cell of mammary gland [purl.obolibrary.org/obo/CL_0002326] with ~53% and fibroblast of breast [purl.obolibrary.org/obo/CL_4006000] with ~16%).

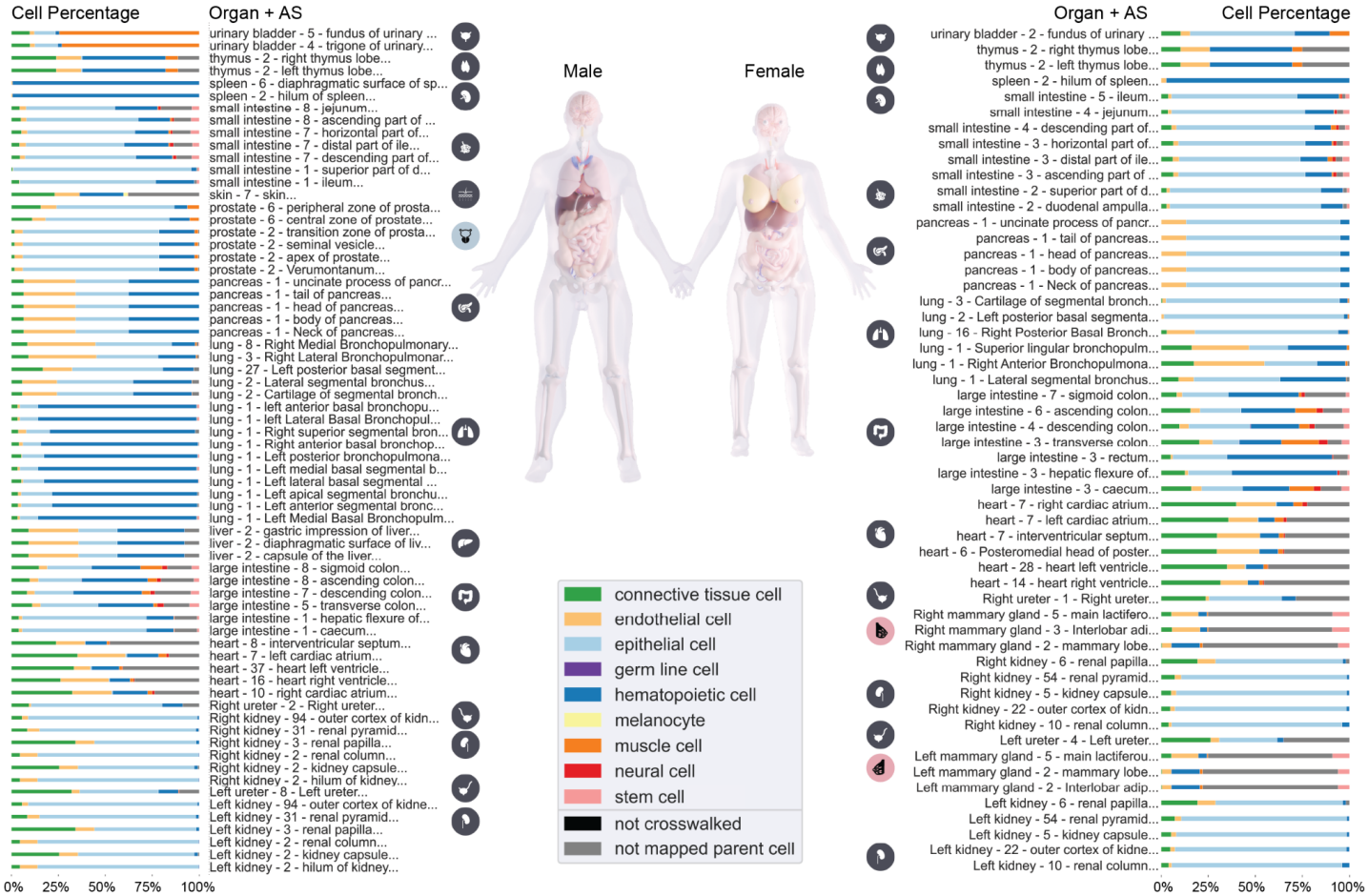

Fig. 1. **CT populations for unique ASs across male and female in HRApop v1.0.** Stacked bar graphs of the percentage of CT identified in ASs, aggregated to higher level CTs from CL, see listing on GitHub[55], shown for male (left) and female (right, placenta was omitted). Male-only organ (prostate) icon is rendered in blue while icons for female-only organs are rendered in pink (mammary gland left and right).

The Opportunity. Using these ASpop based on all the tissue blocks registered into an AS with the RUI, it becomes possible to (1) predict CT populations for 3D volumes in the human body that have sufficient data and (2) predict 3D origin volumes for a tissue block with known ASpop (see **Fig. 2.K** and details in the **Usage Notes** section). In addition, other HRA queries can be run, e.g., a query to retrieve all AS-CT combinations, with sex, tool, CT, and cell percentage is available in the **Usage Notes** section, which also points to Python example code for using this query to get ASpop as well as Dataset and Extraction Site Cell Type Populations (DESpop) from the HRA Knowledge Graph (KG)[22].

Box 1. Key terminology used to explain HRApop. Terms in *italics* have their own entry, either in this Box or in the HRA marker paper[34] or the HRA KG paper[22]. The former already contains definitions for: *3D collision, extraction site, AS, ASCT+B table, biomarker, CT, CTann tool, crosswalk, millitome, polygon mesh, registration set, RUI registered, sc-transcriptomics/sc-proteomics data, tissue block,* and *(weighted) cosine similarity*. The latter contains definitions for: *3D reference object, ds-graph*, and *SPARQL (https://www.w3.org/TR/sparql11-query)*.

- **AS Cell Type Population (ASpop)**: Captures the number of cells per CT for an *AS* given with an ontology ID, donor sex, *AS* label, the *CTann tool* used to assign *CTs*, and a list of datasets from which the data was sourced.

An exemplary *ASpop* is shown on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information/#for-an-as.

- **Cell instance**: Is an occurrence of a unique cell, identified by its bar code or a similar unique ID. In the data products of HRApop v1.0 (see **Data Records**), cells can be annotated by multiple CTann, which leads to potentially multiple occurrences of the same cell.
- **Cell type population**: Is a listing of unique CTs and their counts computed for *ASs, extraction sites*, and datasets. For the latter, mean B expression values are also computed. *CT populations* are computed from *CT* counts in experimental datasets, obtained either via *CTann* in the *DCTA Workflow* (for *sc-transcriptomics* datasets), or via expert/author-provided annotations (*sc-proteomics* datasets). *CT populations* are associated with exactly one *CTann* tool; if a dataset has *CT populations* from multiple *CTann* tools, the *CTann* tool that computed the CT population becomes part of its provenance. *CT populations* are kept inside the *ASpop* graph (for an entire AS) and the *DESpop* (for datasets and extraction sites).
- **Collision detection (bounding box)**: Is a computationally efficient but imprecise method for identifying intersections between two 3D objects in Euclidean space. For each 3D object, the bounding box is a cuboid shape that is calculated that encapsulates all vertices and only this bounding box is used when computing collisions.
- **Collision detection (mesh-based)**: A more precise but computationally more expensive method for identifying the intersection between two 3D objects in Euclidean space based on mesh surfaces, resulting in *AS* tags for extraction sites. To enable mesh-based collision detection, *3D reference objects* need to be pre-processed. Details are provided in the **Methods** section.
- **Corridor**: Describes a 3D volume that encompasses all possible locations for an *extraction site* while maintaining its *intersection percentages* with any *ASs* based on *mesh collision detection*.
- **Criteria C1-4**: Describes four conditions that a dataset must fulfill in order to be used for HRApop construction: (1) it has an extraction site in a 3D reference object, (2) it has a *CT population* via *CTann* or the *sc-proteomics* workflow, (3) it comes from a data portal with QA/QC or have an associated peer-reviewed publication in a scientific journal, and (4) it is from a healthy adult human donor.
- **Dataset and Extraction Site Cell Type Populations (DESpop, formerly "Atlas-Enriched Dataset Graph" in a related HRA publication[34])**: Is a collection of datasets that meet *Criteria C1-4* and are used for HRApop construction. A snippet of the *DESpop* is shown on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information/#for-a-dataset.
- **Download and Cell Type Annotation (DCTA) Workflow**: Is a set of scripts to download H5AD files, execute the Docker (www.docker.com) containers in *HRApop CTann Tool Containers*, and output CT populations and data metadata as input for the RUI2CTpop Workflow.
- **Donor**: Is an individual person, living or deceased, who contributed tissue. Donors have demographic metadata, such as age, body mass index (BMI), race/ethnicity, and sex. All *donors* in HRApop are healthy, human adults.
- **Experimental dataset**: Describes cell by gene matrices (via *CTann tools*) or cell by protein matrices via Organ Mapping Antibody Panels (OMAPs)[56] derived from an organ-specific *tissue block* from a *donor*.
- **HRApop Atlas Data (also called HRApop Data Used in Atlas Construction)**: Is a collection of high-quality experimental datasets that fulfill *Criteria C1-4* that is used in atlas construction. Also part of the *HRApop Atlas* are the *CT populations of* 3D extraction sites of these datasets, plus the ASs with which these extraction sites collide. At the beginning is a programmatically compiled collection of dataset IDs, extraction sites, and donor metadata that forms the input of the *DCTA Workflow*. Dataset IDs can come from three sources: (1) sc-portals with access to sc-transcriptomics H5AD files (from HuBMAP, SenNet, CxG, and GTEx), which are then given to the *DCTA Workflow* as an input, (2) sc-proteomics datasets, and (3) HRA Registrations, see GitHub[57] and **Table S1**, a set of manually curated extraction sites, from which the *RUI2CTpop Workflow* extracts dataset IDs and attempts to match them with *CT populations* from the *DCTA Workflow*. Each HRApop run provides a series of reports (see GitHub[58]).
- **HRApop CTann Tool Containers**: Is a collection of CTann tools containerized with Docker to annotate H5AD files. The DCTA Workflow uses the *HRApop CTann Tool Containers*. They also contain utilities for CTann, e.g., computing gene expressions and identifying the top-n genes.
- **HRApop Prediction Data**: Datasets that fulfill C3-4 *and* have either *CT populations* (fulfilling C2) or *extraction sites* (fulfilling C1) but not both and for which the missing information can be predicted.

- **Intersection percentage or volume**: Is a measure that describes the shared 3D space between two 3D objects (e.g., a *tissue block* and an *AS*) expressed as a percentage of the total volume or as an absolute value in cubic millimeters (mm^3).
- **Mean biomarker expressions**: Are computed for the top 10 genes per CT per dataset, then stored in the *DESpop*.
- **RUI to Compile CTpop (RUI2CTpop) Workflow**: Is a collection of scripts to compile the *HRApop Atlas Data* from the output of the *DCTA Workflow* (see **Fig. 2 and Methods**).

Results

The HRApop effort combines (1) CT populations from sc-transcriptomics and sc-proteomics data made with CTann tools (or via user-assigned CT label), (2) donor metadata, and (3) 3D extraction sites made via the RUI. This involves running two scalable workflows that were optimized to effectively deal with the continuously growing number of datasets. Over the last three years, the processes used to download data, perform cell type annotations, and compute CT populations per dataset (called **DCTA Workflow**) and the workflow that takes CT populations, donor metadata, and 3D extraction sites via the RUI, to compile Cell Type Populations for ASs, datasets, and extraction sites (called **RUI2CTpop Workflow**). These are then published as graphs in the HRA KG[22] for usage in various HRA applications and the HRA Application Programming (API, apps.humanatlas.io/api). Note that CT populations record the number of cells per CT not just for ASs but also for datasets and extraction sites as well as the top-10 biomarkers per CT (datasets only) and their mean expression values (see **Methods** section on how these are computed).

Subsequently, this paper details how these two workflows are used in HRApop construction; describe multiple HRApop usages; and list counts for HRApop v1.0, including raw datasets, extraction sites, experimental data, and data products in the HRA KG, following LOD principles[59] and enabling FAIR[60] data access; then, it details a generalization to sc-proteomics data. This section concludes with limitations and outlook for future work on improving the HRApop Atlas Data presented. In the remainder of the paper, the **Methods** section details implementation of the DCTA and RUI2CTpop Workflows. The **Data Records** section details where data is made available to users. The **Technical Validation** section elaborates on (1) the prevalence of different CTs inside not just the organs but also the ASs of the human body and (2) the number of datasets per organ and AS. Finally, the **Usage Notes** describe how data access is facilitated.

Construction

HRApop is constructed with two automated workflows:

(1) The **Download and Cell Type Annotation (DCTA) Workflow** (see GitHub[61] and **Box 1**) uses scalable, open-source software to programmatically download and annotate H5AD files from four portals: HuBMAP, SenNet, GTEx, and CxG. Cells in these H5AD files are then annotated with containerized CTann tools kept in the **HRApop CTann Tool Containers** (see **GitHub**[62] and **Box 1**); next, the DCTA Workflow crosswalks the resulting CTann labels to CL[52] or Provisional Cell Ontology (PCL)[63], see **Methods** section for details, then compiles donor metadata from each dataset and finally makes the result available as a set of CSV files (for metadata) and JSON files (for CT populations). These can then be used as input for the next workflow.

(2) The **RUI to Compile CTpop (RUI2CTpop) Workflow** (see GitHub[64] and **Box 1**) then computes CT populations for ASs, datasets, and extraction sites. It first identifies all datasets that can be used to construct HRApop based on four Criteria (C1-4, see **Box 1**):
    - **C1:** The dataset has a 3D extraction site by registration with the RUI (see **Methods**), i.e., the spatial position, rotation, and size of the tissue block (from which the dataset was derived) within the HRA reference system is known and AS tags exist.

- **C2:** The dataset has a CT population (see **Box 1**), i.e., number of cells per CT. For sc-transcriptomics data, this means having an associated H5AD file with a cell by gene matrix that can be annotated via Azimuth[48], CellTypist[49,50], and/or popV[51]; for sc-proteomics data, CT populations are obtained via manual segmentation and annotation workflows.
- **C3:** The dataset is of high quality, i.e., it was downloaded from a data portal with built-in quality assurance/quality control (QA/QC) or is associated with a peer-reviewed publication.
- **C4:** The dataset has complete donor metadata, is from a healthy tissue sample, and has an age value greater than 18. That is, the data is from a healthy adult human.

C1 and **C2** are enforced to enable usage of HRApop Atlas Data for HRA User Story (US)[34] #1 (predict CT populations) and US#2 (predict spatial origin of tissue samples). **C3** ensures that atlas data comes from reliable, scientifically backed sources. **C4** is in support of constructing a healthy reference for HRA US#3 (compare reference tissue with aging/diseased tissue). In the future, a disease reference will be built with adjusted criteria.

For datasets that meet **C1-4**, RUI2CTpop outputs the **HRApop Atlas Data** (also called "**HRApop Data Used in Atlas Construction,**" see **Box 1**), a collection of data products that describe high-quality ASpop and DESpop for the HRA.

Datasets that do not meet all four criteria are disregarded, except: Datasets that **fulfill C3-4** and either **C1** or **C2** are added to the **HRApop Prediction Data**, for which HRApop can be used to predict either an extraction site or a CT population (see **Box 1** and **Outlook** section for details as well as **Table S1** for download links).

Fig. 2 shows construction and usage of HRApop, from the data download on the left to the publication and usage of HRApop Atlas Data on the right. It provides an overview of data download and CTann runs via the DCTA Workflow as well as ASpop and DESpop computation via the RUI2CTpop Workflow, see details in the **Methods** section.

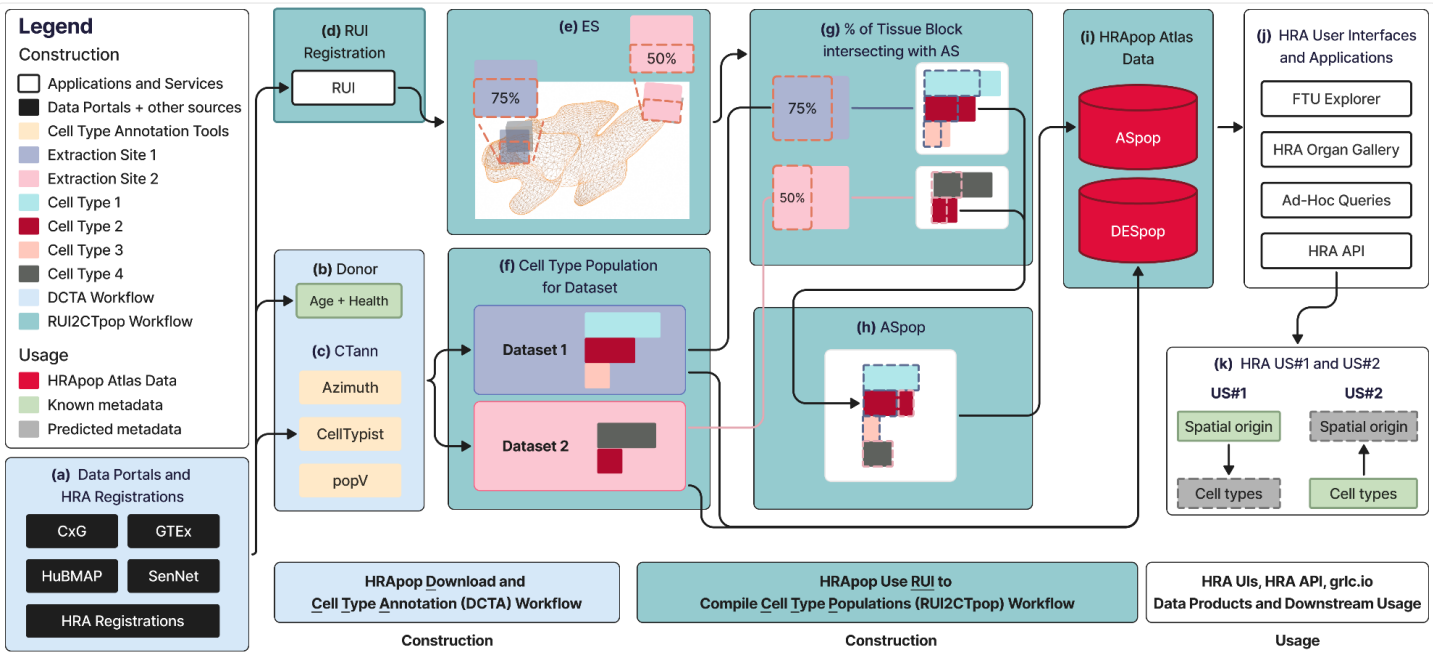

Fig. 2. Construction and usage of HRApop. (a) Data is ingested from four data portals or from HRA Registrations[57]. **(b)** Donor information is extracted from dataset metadata or the extraction site (if from HRA Registrations). **(d)** CT populations are computed using CTann tools if possible. **(a)-(c)** is handled by the **DCTA Workflow**. **(d)** RUI-assigned extraction sites for each dataset are identified. If a dataset has an extraction site, **(e)** intersection percentages (see **Box 1**) of the extraction site with ASs can be computed. **(f)** CT populations for corresponding datasets obtained via one or more CTann

tools can be combined with **(g)** intersection percentages between extraction site and AS, resulting in **(h)** ASpop, which, together with DESpop, are then published as **(i)** HRApop Atlas Data. **(d)-(i)** is handled by the **RUI2CTpop Workflow**. **(j)** illustrates usage of HRApop Atlas Data by HRA applications. **(k)** shows usage for HRA US#1 and US#2.

First, in the **DCTA Workflow** (blue), datasets represented as H5AD files are programmatically downloaded from the four data portals (**Fig. 2.a)** alongside donor metadata (**Fig. 2.b**). Before download, non-human and diseased data is filtered out. Then, each dataset is annotated using all applicable CTann tools (**Fig. 2.c**), resulting in CT populations if the dataset originates from a supported organ (**Fig. 2.f**). Only if a dataset meets all Criteria C1-4 (see **Box 1**), then it is used for HRApop Atlas construction.

Next, in the **RUI2CTpop Workflow** (green), datasets are matched against all existing RUI extraction sites by their ID (**Fig. 2.d)**; the extraction sites hold metadata on organ sex and laterality (left or right) and can come from either an API (such as for HuBMAP or SenNet, see **Methods** section) or the static HRA Registrations[57] (see **Table S1**), a manually curated collection of extraction sites.

For each HRApop Atlas Dataset, i.e., that has been determined to fulfil Criteria C1-4 (see **Box 1**), the 3D ASs that its extraction site collides with are determined via mesh-based collision detection (**Fig. 2.e**, see also **Box 1)**. The example shown is the renal pyramid (purl.obolibrary.org/obo/UBERON_0004200) of the left, male kidney (lod.humanatlas.io/ref-organ/kidney-male-left/latest) with two hypothetical tissue blocks colliding with it at 75% and 50% of the total volume of the extraction site. The number and percentage of CTs that should be in the colliding AS based on the intersection percentage (see **Box 1**) of the extraction site are computed (see **Fig. 2.g**). This is performed for every extraction site that intersects with the AS. The result is an ASpop containing the unique and shared CTs contributed by each colliding extraction site (see **Fig. 2.h** and details in **Methods**).

In cases where there exists just one tissue block (see definition in related HRA publication[34]) that is registered into exactly one 3D AS, this AS is assumed to have the CT population and biomarker expression levels of that tissue block; plus, confidence will be low as only one dataset exists for that AS. Typically, many tissue blocks from different donor demographics (e.g., age, ethnicity) exist per male/female-specific ASs. Often, research teams carefully sample from the very same RUI extraction site (see definition in a related publication[34]) so hundreds of datasets have the very same extraction site. In some cases, a tissue block is registered with more than one AS; here, CT populations are split using the percentage of mesh-based collisions detected (see **Box 1**) to associate cells to ASs. Finally, the ASpop and the DESpop used to generate the ASpop are published as two separate Resource Description Framework (RDF, www.w3.org/RDF) graphs (see **Fig. 2.i**) via the HRA KG[22] at lod.humanatlas.io/graph/hra-pop/latest in support of Linked Open Data (LOD)[59] principles.

Usage

Finally, the data products generated as part of this RUI2CTpop Workflow are then used in various HRA UIs and the HRA API (**Fig. 2.j**), which then makes them available for usage in HRA US#1 and US#2 (**Fig. 2.k**).

HRA KG queries can be run to support diverse HRA applications (see **Fig. 2.j).** Examples are the HRA FTU Explorer (apps.humanatlas.io/ftu-explorer), the HRA Organ Gallery in virtual reality (VR)[65,66], and ad-hoc queries, such as one that provides an overview of all AS-CT combinations, including sex, tool, cell count, and cell percentage, available at apps.humanatlas.io/api/grlc/hra-pop.html#get-/cell_types_in_anatomical_structurescts_per_as. Via the HRA API (apps.humanatlas.io/api), the ASpop and DESpop are then used for the Cell Population Predictor and the Tissue Origin Predictor in support of HRA User Stories[34] #1 and #2 (see **Fig. 2.k** and details in the **Usage Notes** section).

An exemplary ASpop from the kidney and an exemplary snippet of the DESpop are available on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#exemplary-cell-

summaries. A complete listing of all data and code for construction and usage of HRApop is provided in **Table S1.** An overview of all HRA applications that use **HRApop Atlas Data** is provided in **Table S2.**

The **Usage Notes** section introduces a pointer to a series of Jupyter Notebooks that allow users to run the DCTA Workflow on a local H5AD-file, details API queries to access HRApop data programmatically, and details two no-code demo applications (**Cell Population Predictor** and **Tissue Origin Predictor)** provide UIs to implement US#1 and US#2, respectively.

Counts for HRApop v1.0

Raw Datasets

On June 16, 2025, the RUI2CTpop Workflow was run to compile HRApop v1.0. It downloaded 16,293 datasets with 57,911,931 cells from the four sc-portals, which were then sent through a filtering process (see **Methods** section for details). Additionally, as of August 2025, 1,746 tissue blocks from 1,327 different 3D cuboid extraction sites exist across 30 organs and they link to 6,378 tissue datasets.

558 of the 662 datasets in the HRApop Atlas Data are sc-transcriptomics datasets that were annotated using one or more of three major CTann tools **(**Azimuth[48], CellTypist[49,50], popV[51]) covering 11,042,750 unique cells in a total 3D volume of ~12.05 liters (dm^3) with partially intersecting extraction sites in 73 unique ASs in 17 organs (112 ASs and 31 organs if male and female are counted separately). The datasets come from 230 3D extraction sites that cover 54 3D ASs across 17 organs. While the HRApop Atlas focuses on these 558 sc-transcriptomics datasets in 17 organs, the method is generalizable to sc-proteomics (see **Generalization to Spatial Data** section) and to all organs.

Table 1 provides counts of datasets and cells in the HRApop Atlas Data, split by sex, consortium, number of datasets, number of cells, and modality. HuBMAP, SenNet, and GTEx have their own portals. HCA and NHLBI/LungMap[67,68] datasets all come from the CxG Portal. Detailed counts for the HRApop Atlas from the RUI2CTpop Workflow are provided in **Table S3.**

Table 1. HRApop Atlas Datasets. A breakdown of datasets that meet Criteria C1-4 (see **Box 1**) and are used to construct HRApop v1.0.

Sex	Consortium	#Datasets	#Cells	Modality
Female	GTEx	7	47,863	sc_transcriptomics
Male	GTEx	8	70,113	sc_transcriptomics
Female	HCA	63	364,993	sc_transcriptomics
Male	HCA	58	359,273	sc_transcriptomics
Female	HuBMAP	22	900,547	sc_proteomics
Female	HuBMAP	112	2,933,015	sc_transcriptomics

Male	HuBMAP	82	15,676,316	sc_proteomics
Male	HuBMAP	253	6,626,620	sc_transcriptomics
Female	NHLBI/LungMap	1	4,680	sc_transcriptomics
Male	NHLBI/LungMap	2	5,713	sc_transcriptomics
Female	SenNet	18	225,940	sc_transcriptomics
Male	SenNet	36	404,540	sc_transcriptomics
TOTAL		**662**	**27,619,613** sc-transcriptomics: **11,042,750** sc-proteomics: **16,576,863**	

HRA Registrations and Extraction Sites via APIs

For all extraction sites, mesh-based collision detection (see **Box 1**) is used to compute the intersection percentage (see **Box 1**) with 3D ASs, which enables the construction and aggregation of CT populations for these 3D ASs. As a result, extraction sites and ASs in the HRA are enriched with the number of cells per CT from high-quality data. Where possible, for all datasets, extraction sites, and ASs, HRApop provides CT populations from every tool that can annotate the dataset. All extraction sites used in the HRApop Atlas Data are shown in the HRApop-focused Exploration User Interface[24] (EUI) at cns-iu.github.io/hra-cell-type-populations-supporting-information/eui.html. Each extraction site in the HRApop Atlas also has a corridor (see **Box 1** and **Methods** section). By working closely with authors of published, high-quality datasets, as well as tissue providers in HuBMAP and SenNet, RUI is used to generate extraction sites for a growing number of organs and ASs.

Table 2 presents counts for datasets, extraction sites, ASs, and organs based on input data for RUI2CTpop Workflow and the HRApop Atlas.

Table 2. Number of datasets, extraction sites, ASs, and organs covered in HRApop v1.0.

	Datasets	**RUI extraction sites**	**ASs covered**	**Organs covered**
Input data for RUI2CTpop Workflow	16,293	1,132	164	49
HRApop Atlas	662	230	73	17

Experimental Data

The Sankey diagram in **Fig. 3** provides a high-level overview of the input data for the input for RUI2CTpop Workflow along several axes, see below. It can be explored interactively at cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_universe_plotly.html. A version showing only HRApop Atlas Data is

shown in **Fig. S1**, with an interactive version available at cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_atlas_plotly.html.

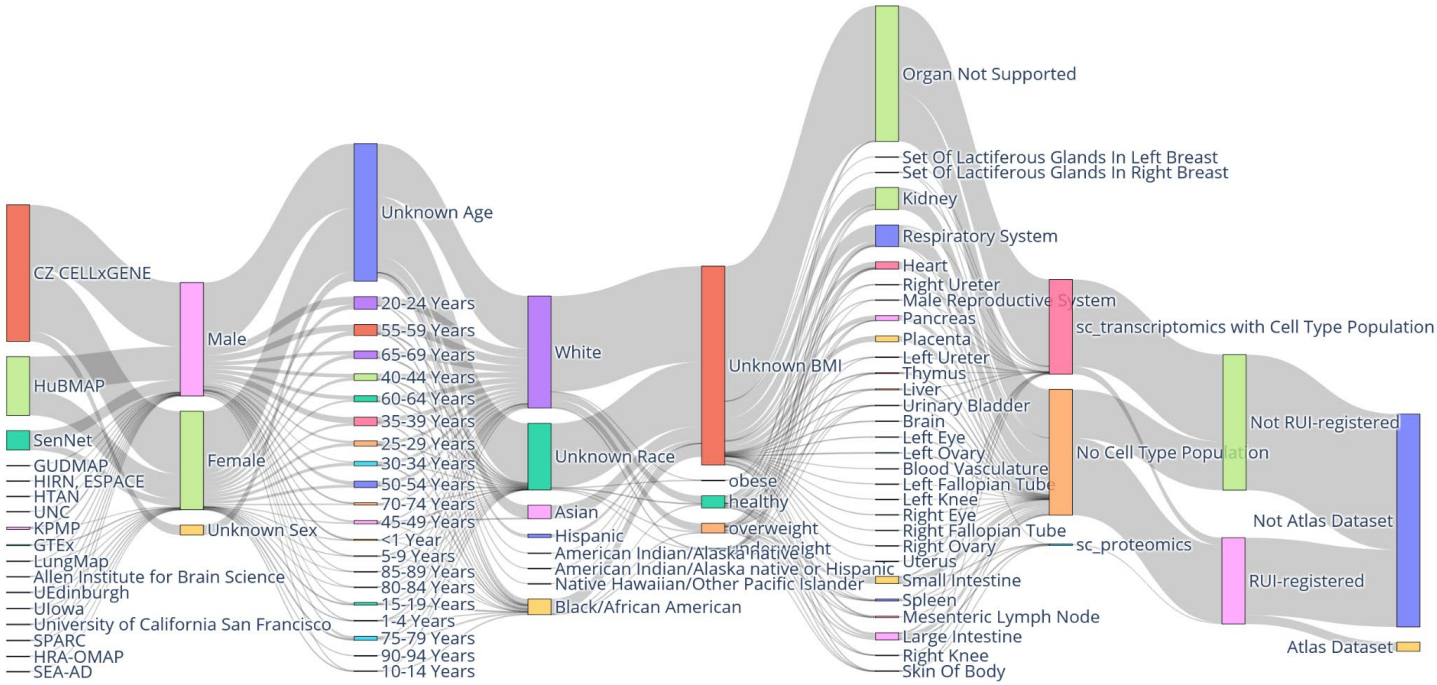

Fig. 3. Sankey diagram of all input data for the input for RUI2CTpop Workflow. Note the Atlas Data node in the bottom right corner, which comprises the 662 HRApop Atlas Datasets in HRApop v1.0.

The Sankey diagram has 9 vertical axes that represent:

Portal/source: Identifies the effort where the data originates. H5AD files were downloaded from HuBMAP, SenNet, GTEx, and CxG. The majority of datasets come from CxG. All other portals/sources were derived from extraction sites. More information about data portals is provided in the **Methods** section.

Donor sex, donor age, *donor race*, *donor body mass index (BMI)*: Describe clinical metadata for the human specimens from which the data was retrieved. Where available, donor age is provided as an integer by HuBMAP and SenNet but as string by CxG (e.g., “61-year-old human stage”); as a result, string literal age values for CxG data were parsed as a number where possible. To enable visualization in a Sankey diagram, all age values were aggregated into bins of five years. For race, the same categories were used as on the HuBMAP Data Portal (portal.hubmapconsortium.org) as of the writing of this paper. BMI values were mapped to categories by brackets defined by the Centers for Disease Control on www.cdc.gov/bmi/adult-calculator/bmi-categories.html.

Organ: Indicates the organ of origin, where “Organ Not Supported” means that there is no matching 3D reference object in the HRA (such as for blood).

CT population: Means that a dataset has either (1) one or multiple CT populations from one or multiple CTann tools (“sc_transcriptomics with Cell Type Population”), (2) no CT population if no CTann tool exists for the dataset (“No Cell Type Population”), or (3) a CT population via sc-proteomics as a generalization of the HRApop workflow (“sc_proteomics with Cell Type Population”).

Extraction site: Indicates whether a dataset has an extraction site via RUI registration or not.

HRApop Atlas Data: Indicates whether the dataset is part of the HRApop Atlas or not.

HRApop in HRA KG

Five data products of HRApop v1.0 are available on Zenodo[69] (see **Data Records** section and **Table S1** for details) and three are available via the HRA KG (lod.humanatlas.io/graph/hra-pop/v1.0):

DESpop (called atlas-enriched-dataset-graph.jsonld on the HRA KG) list CTs with top-10 biomarkers for 662 datasets (558 sc-transcriptomics, 104 sc-proteomics) as well as CTs, mesh collisions, and 230 corridors for 230 3D extraction sites.

ASpop (called atlas-as-cell-summaries.jsonld on the HRA KG) list counts and percentage of CTs for all 73 ASs in 17 organs covered in HRApop v1.0. Additional counts for DESpop and ASpop are provided in **Table 2** above**.**

Corridor GLB files (named after the extraction site on which they are based, e.g., corridors/00087766-0287-467c-9060-b52773db3dce.glb) describe corridors, see **Methods** section. The HRA KG makes 1,189 available, including the 230 corridors for the 230 extraction sites in DESpop, plus 959 corridors for extraction sites not in the HRApop Atlas. Their total size is 202 MB. When compressed and placed on Zenodo[69] as a ZIP file, they are 99.8 MB large.

Generalization to Spatial Data

In HRApop v1.0, 104 sc-proteomics datasets contribute 16,576,863 cells, which brings the total number of cells to 27,619,613. They are associated with high-quality publications[34,70–74] and use protein and antibody-based modalities such as Cyclic Immunofluorescence (CyCIF)[75], Cell DIVE[76–78], and co-detection by indexing (CODEX)[79] to identify proteins and quantify their expressions in a tissue in situ; there are plans to integrate iterative bleaching extends multiplexity (IBEX)[80,81] datasets in the future. While this paper does not focus on sc-proteomics data, it presents a curated collection of sc-proteomics datasets as a generalized use case for HRApop. Spatial proteomics has received significant interest from the scientific community in recent years[82], which has led to increased high-quality data generation. If integrated in HRApop, it enables creating CT populations with a preserved spatial context for each cell (which is lost in sc-transcriptomics datasets, although some recent work has attempted recoveries for specific assay types and organs[47]).

The DCTA Workflow outputs CT populations and metadata for all datasets, both sc-transcriptomics and sc-proteomics. While sc-transcriptomics datasets are run through at least one CTann tool, sc-proteomics datasets have CTs assigned by their expert providers, who share CT populations in the form of CSV files, see code in **Table S1**. The DCTA Workflow systematically goes through a list of sc-proteomics datasets (available on GitHub[83]) and transforms the files with nodes for each cell in each dataset (CSV) into a CT population (JSON) as an input for the RUI2CTpop Workflow (see GitHub[84]). In the future, the DCTA Workflow will be extended to (1) handle other generalized use cases that cannot be annotated with CTann tools and (2) run more CTann tools over new and existing HRApop datasets for additional CT populations.

To make crosswalks for sc-proteomics data, CT labels were shared by contributors of high-quality datasets; crosswalks for each dataset were manually created for that publication effort and made available at lod.humanatlas.io/ctann/vccf/latest. Unmapped CTs are on GitHub[85].

Limitations

HRApop v1.0 comes with a number of limitations:

Dataset duplication across data portals: Some datasets (e.g., by KPMP) are available via multiple data portals (e.g., HuBMAP and CxG). However, no dataset should be used twice for HRApop construction. Data duplication detection is difficult as different versions of the data might exist and different metadata might be associated with each version. For example, there might be a dataset submitted with the very first paper submission and linked to a preprint on a preprint server, a slightly expanded dataset associated with a revised and later preprint version, and a final version of the dataset linked to a peer-reviewed published paper; paper title and authors might also change in the process making deduplication challenging.

CTann tools are trained on underspecified data: CTann tools are trained on high-quality reference datasets that might not be separated by sex or even donor and for which RUI registration information is not available; i.e., it is unknown from which of the diverse ASs within an organ a tissue that was used to train the CTann tool originated. HRApop, however, utilizes existing CTann tools to compile ASpop that are specific to the diverse ASs and tailored to the male and female human body. As more RUI registered sc-datasets become available, CTann tool developers might like to use this additional tissue origin information to optimize CTann.

Missing assay type information needed for batch correction: At present, the ds-graph HRA Digital Object type[22], which captures datasets and their extraction sites plus donor metadata, only contains assay type metadata as free text string rather than not ontology terms (e.g., 0x 3' v1-3, 10x scATAC-seq, MERFISH, Smart-seq, all listed on the CxG portal). HRA KG remodeling is underway to provide look-up tables for assay types to ontology terms for enabling more systematic queries.

Intersecting 3D reference objects: The 1,283 3D ASs are supposed to have no overlap with each so that CT populations are specific to one, not multiple intersecting ASs. However, in HRA v2.3, 18 ASs in 17 organs (e.g., two in the female heart) have intersections with each other and also have extraction sites in them (labeled “TB3” in **Fig. S2**). The AS-AS pairs with the most (here three) extraction sites that collide with both ASs are the ‘outer cortex of kidney’ and the ‘renal pyramid A’ in the male, left kidney and the ‘kidney capsule’ and the ‘outer cortex of kidney’ in the female, left kidney. Intersections will be corrected and tissue blocks will be re-registered for an upcoming HRA release.

Limited coverage of organs and ASs: The HRApop effort aims to enrich all 1,283 3D ASs in 73 3D reference objects of the entire HRA v2.3 with CT populations from high-quality datasets. However, at present, only 73 3D ASs (112 if male and female are counted separately) in 17 organs (31 if male and female are counted separately) have data to compute AS-specific CT populations. The **Technical Validation** section expands on current HRApop coverage of organs and ASs.

Prediction accuracy: Due to the limited number of spatially registered datasets in HRApop, the ability to use HRApop to predict CT populations in support of HRA US#1 and the spatial origin of tissue samples in support of HRA US#2 is limited. About 1,000 more datasets with more than 50 million cells will be added in the next HRA release, see **Outlook** section. Moreover, as shown in the **Technical Validation** section, there exist discrepancies in coverage between male and female organs and CTann tools, where datasets from male donors and specific organs (heart, large/small intestine, lung) are overrepresented.

Outlook

Regular HRApop releases are planned and scheduled, with data products made available via the HRA KG[22] (lod.humanatlas.io/graph/hra-pop) though the HRA Content Delivery Network (CDN)[22] and Zenodo[69]. To improve coverage and quality of HRApop, the following steps are planned:

Increase number of RUI registered datasets: As of August 2025, the HuBMAP and SenNet portals list 842 RUI registered datasets that are currently in QA/QC status but will be published soon. Outreach efforts to authors of peer-reviewed, published papers are ongoing to register their data for use in the DCTA Workflow in support of HRApop Atlas Data construction. The HRApop effort will also integrate data from Tabula Sapiens[5], KPMP, the Helmsley Gut Cell Atlas[86], and the Deeply Integrated human Single-Cell Omics (DISCO) database[87], which has a total of 19,717 datasets, out of which 32% of the total data is from a healthy human body, across 159 unique ASs.

Scale up tissue registration via millitomes: A millitome[34] (from Latin *mille*, meaning “thousand,” as in millimeter, and the Greek *temnein*, meaning "to cut") is a device designed to hold a freshly procured organ and facilitate cutting it into many small tissue blocks of well defined size for usage in sc-analysis and HRA construction. It is used to produce uniformly sized slices or cubes of tissue material that can be registered to 3D reference objects. Using a millitome improves efficiency by enabling consistent, high-throughput sampling.

Improve generalizability: The 104 sc-proteomics datasets in HRApop v1.0 were presented as a generalization from sc-transcriptomics datasets. In the future, and in synergy with HRA Vasculature Common Coordinate Framework (VCCF) construction efforts[72,88], more CODEX[79] datasets and new modalities, such as the Spatial Multiomics Single-Cell Imaging platform CosMx[89], will be added to HRApop.

Increase number of CTann tools used to enable more benchmarking: Currently, HRApop uses three well-established CTann tools backed by scientific publications describing the methods, results, and validations for each tool. Results are presented as CT populations by CTann tool—users can pick their favorite tool and data or perform comparisons and benchmarks between CTann tools[90]. Future HRA releases will feature additional CTann tools such as FR-Match[91]. This will further enable CTann tool comparisons and benchmarking for the same source data.

Compare HRApop with similar efforts using the same source data: Plans exist to compare HRApop with other efforts for computing CT populations. CxG data is heavily used by the sc-community. As a result, benchmarks exist that enable the validation of HRApop CTann results against those obtained by other teams. For example, the CuratedAtlasQueryR package (stemangiola.github.io/CuratedAtlasQueryR) on Bioconductor[92,93] harmonized 12,981 sc-RNA sequencing samples with 29 million curated cells from 45 anatomical sites.

Return confidence scores in Cell Population Predictor: At present, the Cell Population Predictor (see **Usage Notes** section) returns a CT population for the provided extraction site, including the number, percentage, label, and ontology ID of all CTs separated by CTann tool. Updates to the HRA API are planned to enable this UI to return confidence scores for each CT, like the ones in the cell instances (see **Box 1**) for sc-transcriptomics datasets in HRApop v1.0 on Zenodo[69].

Validate HRApop for HRA US #1 and #2[34]: The Cell Population Predictor and the Tissue Origin Predictor are proof-of-concept UIs with underlying HRA API endpoints that enable no-code predictions, see **Fig. 4**. Current work aims to validate and improve these predictions for 617 unique extraction sites (US#1) and 5,453 unique datasets (US#2) in the HRApop Prediction Data that is part of HRApop v1.0.

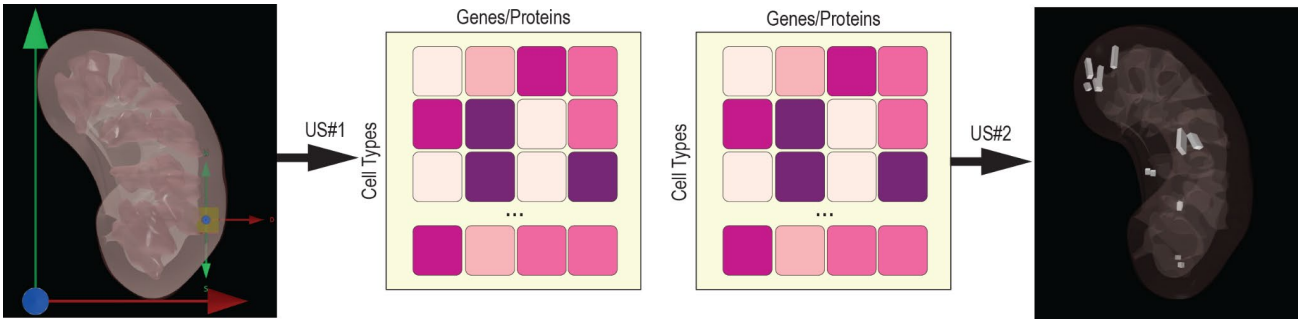

Fig. 4. A simplified illustration of HRA US#1 (left) and US#2 (right). Shown for US#1 is a collection of predicted extraction sites.

Decrease run time for HRApop code: For HRApop v1.0, the DCTA Workflow started on Thu, May 15, 2025, ran for about 10 days, and finished on Sunday, May 25, 2025. It averaged 87.63 dataset annotation runs per hour. Annotations took about 8.59 days to finish. This long runtime is primarily due to the complexity of the annotation algorithms to cover 10s of millions of cells. Targeted optimization of the algorithms and workflows combined with more hardware resources and re-using annotations from prior runs will be required to reduce runtime. Work is underway to save annotations between runs to skip the re-annotation step and work faster. After a 22-day QA phase, the RUI2CTpop Workflow started on June 16 at 5:55:27 PM EDT and finished about four hours later at 10:07:11 PM EDT the same day. A full log is available in **Table S1**. In the future, the run time for the DCTA Workflow will be decreased by using high performance computing (HPC) in the form of Big Red at Indiana University (kb.iu.edu/d/brcc). Also, the crosswalking will be moved to the RUI2CTpop Workflow, which will decrease runtime and improve modularity of both workflows.

Methods

Estimates of Number of CTs in the Human Body

While no consensus exists on the number of CTs within the human body, estimates range from 400 major CTs[3–6] to 3,358 CTs[7], depending on the criteria used to determine what constitutes a CT. As an example, in the retina, a major class of retinal neurons is the amacrine cell (purl.obolibrary.org/obo/CL_0000561). This cell can be subdivided in multiple ways. At a broad level, it is often classified into GABAergic (purl.obolibrary.org/obo/CL_4030027) and glycinergic (purl.obolibrary.org/obo/CL_4030028) types (two categories). However, based on morphology, researchers have identified 25 distinct amacrine CTs[94]. At the level of sc-transcriptomics, the Human Retina Cell Atlas (HRCA)[95] has reported 123 CTs in the retina, including 73 molecularly distinct amacrine CTs. As a result, depending on the resolution—functional class, morphology, or transcriptomic profile—one might count one, two, 25, or 73 different amacrine CTs. In the case of the human brain, sampling more than three million nuclei from approximately 100 dissections across the forebrain, midbrain, and hindbrain, 461 clusters and 3,313 subclusters (granular CTs) organized largely according to developmental origins were identified[96]. Combining CTs in the Human Lung Cell Atlas[97] and the CellRef atlas from the LungMAP consortium[98] revealed 68 distinct CTs in the human lung and nasal cavity[90]. As of August 2025, CL[52] contains 3,358 classes, but this includes many non-human terms as well as grouping classes—i.e., internal nodes in the hierarchy that do not correspond to distinct, terminal CTs. When limiting the count to leaf-level human CTs, the number is on the order of 2,500, but this does not reflect many of the novel CTs that have been recently defined using sc-technologies. Further, about 2,000 CTs[7] in CL are connected to ASs in the cross-species anatomy ontology Uberon[99,100] via 'part of' relationships. As a result, 400 total CTs might be a clear underestimation. A more plausible estimate is closer to 1,900, though the exact number depends on the criteria used to define and distinguish CTs. Assuming that there are 78 major organs in the human body of varying size and cellular complexity, with most organs averaging between 50 and 120 transcriptomic CTs,

aside from the brain with 3,000-5,000, there might be close to 10,000 CTs in adult mammalian organisms, depending on the criteria for distinguishing CTs.

Data Used

Data Portals

HuBMAP: As of August 2025, the HuBMAP Data Portal (portal.hubmapconsortium.org) lists 4,742 public datasets from 27 organs. These datasets are ingested by tissue providers through a UI (ingest.hubmapconsortium.org) where they can enter donors, organs, samples/tissue blocks, tissue sections, etc. Relationships between these entities are organized in a provenance hierarchy where a donor and organ are needed so that tissue samples can be organized based on diverse tissue sample types. APIs enable users to access entities programmatically (docs.hubmapconsortium.org/apis.html). Both published and unpublished datasets exist on the HuBMAP and SenNet portals (see below). Published datasets have been sent through a series of QA/QC processes. Unpublished datasets are only accessible with authentication. Only published datasets are used for HRApop construction.

SenNet: The SenNet Data Portal (data.sennetconsortium.org/search) uses a similar infrastructure as the HuBMAP Data Portal. It features human and murine datasets. As of August 2025, 1,712 human datasets from 903 donors are publicly available. Like for HuBMAP, APIs provide programmatic access to datasets, donors, organs, etc. (docs.sennetconsortium.org/apis).

GTEx: The GTEx Portal (www.gtexportal.org/home/downloads/adult-gtex) hosts the adult GTEx data and resources and provides open access to, e.g., expression quantitative trait loci (eQTLs), and protected access, e.g., to limited donor phenotypes as well as de-identified donor data for sequencing. For GTEx sc-data, one H5AD file with data for all 8 organs plus donor metadata was downloaded (storage.googleapis.com/adult-gtex/single-cell/v9/snrna-seq-data/GTEx_8_tissues_snRNAseq_atlas_071421.public_obs.h5ad).

CxG: The CxG Portal (cellxgene.cziscience.com/collections) provides access to both primary and secondary datasets. Primary datasets contain the raw or minimally processed data while secondary datasets are curated and normalized. The DCTA Workflow retrieves the secondary datasets when possible unless they contain fewer data points than the primary datasets. As of August 2025, there are 1,183 collections with 791 healthy adult donors.

Cell Counts

Cell counts represent the unprocessed number of RNA transcripts detected for each gene in each cell. To ensure data integrity and consistency, raw cell counts are used wherever available during CTann. For HuBMAP and SenNet datasets, these are obtained from the “counts” layer of the H5AD file, which contains integer counts per gene and cell, while for GTEx and CxG datasets, raw counts are obtained from the “raw.X” attribute of the H5AD file, which stores original and unnormalized counts for each cell and gene.

Crosswalks

To enable comparisons between CTs assigned by different CTann tools (and sc-proteomics data, which use human-assigned CTs), CT labels need to be crosswalked to CTs in the anatomical structures, cell types, plus biomarkers (ASCT+B) tables[35] using CL or PCL terms. Crosswalks for each CTann tool are curated manually by experts and are published at lod.humanatlas.io/ctann; the underlying HRA Digital Object type is described in a related publication[22]. Since not all CTs have an exact match to a CL term, the Simple Knowledge Organization System (SKOS, www.w3.org/2004/02/skos)[101,102] is used to indicate if the mapping was done to a

term with an exact match (skos:exactMatch) or if they had to be mapped to a more general class (skos:narrowMatch). The crosswalks published in HRA v2.3 link 1,615 annotation labels and 1,909 annotation IDs from Azimuth, CellTypist, popV, and sc-proteomics datasets (author-assigned) to 885 CL labels and 495 CL IDs for 36 organs. 1,923 mappings are exact matches and 683 are narrow matches.

Fig. 1 uses CTann crosswalks to harmonize and compare datasets across portals. The resulting CT typology has 201 low-level CTs that were aggregated into 9 high-level CTs—a subset of broad terms curated by CL editors that covers the majority of CTs with minimal overlap. Additional mappings exist to 19 out of 24 medium-level, more granular CTs—a subset of terms that, while still broad, provides a more detailed classification—via CL IDs and labels; this mapping is available in **Table S1**. The mapping applied in **Fig. 1** uses the 9 CTs belonging to the upper slim of high-level CTs directly from CL (see GitHub[103]); note that CTs from PCL are included in this slim as well. 26 CTs were classified under more than one category in CL; for these, one of the classes was chosen reflecting expected biological grouping (see **Table S4**). If a CT is marked as "no mapped parent cell," the CT term (already crosswalked to CL or PCL) is not a subclass of a top-level CT in the top-n CTs provided by CL. If a CT is marked as "not crosswalked," this means that the CT label assigned by the CTann tool was not associated with a CL or PCL ID. A full report of cells that are not crosswalked while constructing HRApop but that have been crosswalked semi-manually for **Fig. 1** is labeled "unmapped-cell-ids" and is available on GitHub[104]. No general CT could be identified for 617,000 cells that are associated with 10 CTs (called "not mapped parent cell" and shown in gray) and 9 cells are associated with 1 CT in the small and large intestine that are not covered in existing ontologies (called "not crosswalked" and rendered in black).

The DCTA Workflow applies crosswalks after CT annotations are done; CL labels and IDs are used when computing CT populations for datasets, extraction sites, and ASs. Exactly 154 unique cell labels could not be crosswalked to CL and a report of modality, CTann tool (if applicable), organ, and CT label for these in **Table S1**.

Existing Code

CTann Tools

The three CTann tools in HRApop v1.0 (Azimuth[48], CellTypist[49,50], popV[51]) are containerized with Docker in the HRApop CTann Tool Containers on GitHub[62]. The Dockerfile lists operating system requirements, basic setup, and dependencies so the package can be run in the cloud or on a local machine with consistent outputs. This enables the DCTA Workflow to execute them locally. These containers define a dataset handler interface (see code on GitHub[105]) that specifies requirements for every new piece of code that downloads H5AD files from a portal. Docker is used to package each CTann tool along with all of its dependencies into portable containers that can be deployed across different environments. Apptainer (apptainer.org), is used when running HRApop code on HPC clusters. An example for Azimuth is provided on GitHub[106]. Each container has a context file that specifies which models need to be downloaded, such as this one[107] for Azimuth. Similarly, extracting summaries, computing gene expressions, and crosswalking are also packaged as Docker files. Both Docker and Apptainer/Singularity are supported for running Common Workflow Language (CWL, https://www.commonwl.org) workflows on a Linux cluster for HRApop v1.0. **Table S5** lists the name, version number, code base, models used, and requirements for each tool.

Mean Biomarker Expressions for sc-transcriptomics Data

Mean biomarker expressions are captured in the DESpop and provided for each CT per dataset. To generate these values, scanpy[108], numpy[109], and anndata[110] are used. Specifically, scanpy's *rank_gene_groups()* method (scanpy.readthedocs.io/en/stable/generated/scanpy.tl.rank_genes_groups.html) is applied to perform differential expression analysis between CTs. As part of this analysis, this method calculates the mean

expression of each gene within a target CT, as well as its expression in the rest of the dataset. These values are used to identify and rank marker genes, and the corresponding mean expressions are recorded for the top-n genes, where n is defined by the user. However, it is important to note that this method does not compute mean expression values for all genes across all CTs—only for the most differentially expressed ones. To ensure consistency in gene naming across datasets, gene identifiers are normalized using a lookup table, which maps Ensembl IDs from Release 111[111] (www.ensembl.org/index.html) to HGNC-approved symbols from version v2023-09-18[112] (www.genenames.org). This normalization helps maintain interoperability and accuracy when comparing gene expression data across datasets and tools.

Tissue Registration

Registering tissue datasets inside the 73 3D reference objects in the HRA is possible via the RUI[24], which generates 3D extraction sites; it is available as a standalone tool at apps.humanatlas.io/rui but also embedded into the ingest pipelines of HuBMAP and SenNet. The registration process consists of three main phases: assignment, enrichment, and validation.
Initially, the registration coordinator contacts subject matter experts with knowledge of the spatial information associated with tissue samples. Experts can then use the RUI to submit tissue block spatial information themselves or collaborate with the coordinator, who facilitates the submission process with their input. The RUI records spatial information by creating extraction sites inside 3D reference objects. Additionally, it uses a mesh-based collision detection to annotate the extraction with AS tags (see **Box 1** and **Mesh-Based Collision Detection** section). By the end of this phase, all tissue samples should have an assigned extraction site. These workflows are further detailed in SOPs[26–28].

Once extraction sites have been assigned, enrichment begins. The registration coordinator uses a location processor tool to enhance spatial information with de-identified donor metadata (e.g., sex, age, BMI) and publication metadata (e.g., DOI, authors, publication year). This combined dataset forms a registration set[34], which is assigned a unique ID for future reference. The code for the processor is accessible via GitHub[113], as are existing stand-alone HRA Registrations[57].

Finally, in the validation phase, the expert is asked to review the registration set for accuracy and completeness, requesting revisions if necessary. This process uses a customized instance of the EUI, which the expert uses to evaluate sample locations, metadata accuracy, and AS tags from mesh-based collision detection. Once validated, the registration set is finalized and added to the general EUI (apps.humanatlas.io/eui). This concludes the spatial registration process. An overview of all EUIs for registration sets is available on GitHub[114].

Mesh-Based Collision Detection

To enable more precise collision detection between tissue blocks and ASs, a library for mesh-based collision detection (see **Box 1**) was created; it is called the HRA Mesh Collision API[115]. Given an extraction site, this HTTP service returns a list of mesh collisions with ASs and metadata. The 3D geometry-based tissue block annotation code includes: (1) a C++ library for the HTTP service for collision detection and intersection volume (see **Box 1**) computation between extraction sites and ASs, (2) a C++ library for checking manifoldness and closedness of meshes as well as hole-filling for unclosed meshes, and (3) a Python library for converting GLB files to Object File Format (OFF) files, used as the underlying 3D model format for collision detection. Code repository, URL to deployed API, and exemplary API response are available in **Table S1**.

Weighted Cosine Similarity

A collection of functions to compute and use weighted cosine similarities in the RUI2CTpop Workflow is available on GitHub[116]. The script uses math.js (mathjs.org) for access to implementation for the dot product

and norm between two vectors (mathjs.org/docs/reference/functions/dot.html and mathjs.org/docs/reference/functions/norm.html).

New Code

HRApop CTann Tool Containers

These are a collection of scripts (see GitHub[62]) for annotating H5AD files using Azimuth[48], CellTypist[49,50], and popV[51]. They wrap CTann tools as Docker containers for CWL workflows. This allows each tool to be executed consistently regardless of technology needed, e.g., R vs Python. Default settings are used for each tool where possible. A complete list of settings for each tool is provided in **Table S5.**

DCTA Workflow

HRApop datasets download and CTann is handled by the **DCTA Workflow** (see GitHub[61]). It uses the tools in **HRApop CTann Tool Containers**, to download data from multiple portals after ensuring they are from healthy, human donors, runs applicable CTann tools, analyzes gene expressions to identify top genes, crosswalks CT labels from CTann to ASCT+B tables[35] using crosswalks, assembles donor metadata, and outputs summarized results for downstream use. It is runnable as a CWL workflow. The CWL runner is written in Python. The DCTA Workflow outputs CT populations and metadata for all annotated sc-transcriptomics and sc-proteomics datasets (see **Table S1**). They are then copied to the input GitHub repository[117] for the RUI2CTpop Workflow (see below) for further processing. The **Usage Notes** section points to example code that runs the DCTA Workflow locally on a user-provided H5AD file.

Download

To download H5AD files, the DCTA Workflow constructs a series of jobs to execute. An organ mapping provides crosswalks between organ code names on the portals to Uberon IDs and labels on HuBMAP and SenNet (see this GitHub commit[118]) as well as GTEx (see this GitHub commit[119]). For CxG, no mapping is needed as the metadata contains UBERON IDs already. To retrieve donor metadata across portals, each portal is queried according to their APIs, relevant information is extracted and saved in a harmonized format, see donor field at this GitHub commit[120]. Implementations to extract donor metadata from the different portals are also available, e.g., for SenNet (see this GitHub commit[121]) and CxG (see GitHub[122]).

The DCTA Workflow extracts metadata needed for constructing ds-graphs[22] (age, sex, BMI, assay type) from the individual portal APIs and saves it as simple JSON files. The H5AD files are downloaded locally into a raw data folder in the DCTA Workflow repository, or onto a file system on a HPC system.

Splitting and Re-assembling H5AD Files for GTEx and CxG data

In the data model of HuBMAP and SenNet, donors, organs, tissue blocks, tissue sections, and datasets are modeled as individual entities, where each dataset belongs to exactly one donor. This means that an H5AD file from HuBMAP or SenNet contains data for exactly one donor. On CxG, on the other hand, H5AD files contain multiple donors; to make the two data models work together, H5AD files from CxG need to be split into new H5AD files by donor and organ in the DCTA Workflow. The respective script[123] is in Python because it needs pandas, a foundational library for data manipulation and analysis (pandas.pydata.org), and anndata for handling annotated data matrices (https://anndata.readthedocs.io/en/stable). This is done to combine donor-organ combinations across assets into new H5AD files. Extracted donor metadata fields are shown in the harmonized donor metadata format described above.

Datasets with too few cells

Datasets with fewer than 100 cells are filtered out by the DCTA Workflow. While their H5AD files are downloaded, no CTann is run over them and no CT population is output.

RUI2CTpop Workflow

The **RUI2CTpop Workflow** (see GitHub[64]) performs spatial annotation and CT population computation with input files provided by the DCTA Workflow. It sources extraction sites via the HuBMAP and SenNet APIs through HRA API queries at apps.humanatlas.io/api#get-/ds-graph/hubmap, apps.humanatlas.io/api#get-/ds-graph/sennet, and apps.humanatlas.io/api#get-/ds-graph/gtex, with the underlying queries at github.com/x-atlas-consortia/hra-api/blob/main/src/library/ds-graph/operations/hubmap.js, github.com/x-atlas-consortia/hra-api/blob/main/src/library/ds-graph/operations/sennet.js, and github.com/x-atlas-consortia/hra-api/blob/main/src/library/ds-graph/operations/gtex.js. A listing of all extraction sites that are sources is available at https://github.com/x-atlas-consortia/hra-pop/blob/main/input-data/v1.0/config.sh.

For the RUI2CTpop Workflow to function, the DCTA Workflow (see above) provides CT populations and dataset metadata, then those files are copied over to the input folder for a new RUI2CTpop run (see GitHub[84]). Scripts running over these input files during RUI2CTpop are on GitHub[124]. Output data from **RUI2CTpop** is provided on GitHub[125].

It processes Criteria C1-4 (including a check for donor age to ensure only data from adult humans is used, see **Box 1**), then gathers extraction sites, CT populations, and donor metadata, including publications. If a dataset has no metadata for age or sex, it is not used for atlas construction. Note that GTEx provides only age ranges, not values, but the data comes from adult donors. It also uses mesh collisions and corridor code to compute corridors (see **Corridors** section) and to build the ASpop and the DESpop (see both definitions in **Box 1,** links in **Table S1**). Exemplary CT populations for a dataset, an extraction site, and an AS are shown at cns-iu.github.io/hra-cell-type-populations-supporting-information/#exemplary-cell-type-populations. Note that in all three cases, the CTann tool(s) are indicated by the annotation_method field. The RUI2CTpop Workflows also contains scripts and SPARQL queries to construct data products for HRApop in the form of CSV reports[58] to analyze, visualize, validate, and use HRApop data (see link in **Table S1**).

Corridors

For each extraction site with a CT population, a 3D volume of likely origin within the organ is computed, given the biomolecular make-up of the tissue block as represented by its CT population. The result is a complex corridor, i.e., a combined representation for all possible locations, compiled via a shrink-wrap algorithm[126].

A corridor represents the complete set of spatial positions where an extraction site could plausibly be located, while maintaining its observed intersection ratios with neighboring ASs. Each extraction site is uniquely associated with one such corridor. Corridors are GLB files (see **Data Records**). In HRApop usage, given a CT population for a dataset of unknown spatial origin, HRApop can be used to predict its spatial origin (HRA US#2). The spatial origin can be an entire AS if it has the same or a similar CT population (measured using weighted cosine similarity) or the extraction site of a tissue block with the most similar CT population (and its corresponding corridor with the same percentages of multiple ASs).

To generate complex 3D corridors given an extraction site, a C++ library with an HTTP service for the 3D Corridor Generation API[127] given an extraction site made with the RUI was created. A corridor is computed by sending an extraction site to the HRA API at apps.humanatlas.io/api/#post-/v1/corridor (see also **Table S1**) to get the mesh-based AS. From there, three cases are possible: (1) If the tissue block collides with **only one AS**, the entire AS is returned as a corridor; (2) if the tissue block collides with **two ASs,** a filter-search algorithm is

used to efficiently compute all the possible locations before applying the shrink-wrap algorithm to generate a complex corridor. The filter-refine paradigm[128] is widely used in computationally intensive tasks where infeasible solutions are filtered out from a list of candidates. Next, more viable candidates are examined with respect to their exact geometry to generate exact answers in a refinement step. Inspired by the filter-refine paradigm, a filter-search algorithm is proposed to derive complex corridors. Finally, (3) if a tissue block collides with **three or more ASs,** it is fixed in place, in which case it corresponds exactly to the extraction site.

A listing of all corridors as well as the code, endpoint, and documentation for the 3D Corridor Generation API is available in **Table S1**. An exemplary corridor belonging to the extraction site with ID 1cbd9283-2d58-4a2d-88fe-effb18c3f14f from the head of the female pancreas that can be inspected in 3D is available on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#exemplary-corridor.

Tradeoff between Step Size in Search Stage and Precision of Corridor

Since a sliding window approach is used to search feasible locations, the configuration of the step size is essential. It determines how big a move is made in the search stage (see **Table 3**). If a large step size is set, locations with the exact intersection volume (see **Box 1**) with the given extraction site may be skipped. Conversely, if a small step size is set, the computation cost may become overwhelming. Further, the intersection volume from the mesh-based collision detection is returned as a float; thus, if only the exact value is matched, feasible locations can be few. In order to compute corridors with both high precision and reasonable computation overhead, the tolerance for feasible locations can be adjusted to, e.g., 0.1, which means the difference between the intersection volume of the feasible locations and the true intersection volume cannot exceed 10% of the true intersection volume.

Table 3. Filter and search stage for pre-computing corridors.

1	**Filter stage**	Each collided mesh is approximated by their minimum bounding boxes. Then, the area $\boldsymbol{\Omega}$ where a fixed-size, axis-aligned tissue block can be put is computed so that it can intersect with all the minimum bounding boxes of meshes.
2	**Search stage**	A brute force search algorithm is applied by specifying the step size. A sliding window approach is used to move the fixed-size tissue block within the searching area $\boldsymbol{\Omega}$ computed in the filter stage.

Data Records

Five major HRApop data products are available for download on Zenodo[69]: Atlas ASpop and the DESpop (both as JavaScript Object Notation for Linked Data format, or JSON-LD, see json-ld.org), as well as cell instances (see **Box 1**) with confidence scores and cell instances with top 10k genes (both as compressed CSV). Additionally, corridors are available as a compressed folder with GLB files.

The Atlas ASpop, DESpop, and corridors are also available via the HRA KG at purl.humanatlas.io/graph/hra-pop/v1.0 and on GitHub[125], as well as in the form of canned SPARQL queries at apps.humanatlas.io/api/grlc/hra-pop.html. Finally, links to all data are provided in **Table S1,** as is a complete listing of all code to construct and use HRApop.

Examples for usage of this HRApop Atlas Data are available on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#usage-examples. The **Usage Notes** section details how to access HRApop data via Jupyter Notebooks.

Technical Validation

Heatmaps

While some CTs exist across organs (e.g., macrophages), most CTs are highly specialized to deliver well-defined organ-specific functions in ASs. To demonstrate that ASpop varies not only by organ but also by AS, four heatmaps were made (one per CTann tools plus sc-proteomics, see **Fig. C1** and **Fig. C2** on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#figures. Each heatmap lists CT labels on the x-axis and organ plus AS labels on the y-axis. Table field color represents the scaled mean value (z-score, see **Equation 1**) for the percentage of CTs identified in each AS.
For each heatmap, data from a CTann tool is selected and processed to calculate the average CT percentage associated with all the ASs in an organ. The results are transformed from a data frame into a matrix (CTs by organ plus AS concatenated into a combined label), where each matrix cell represents the average CT percentage measured for CT and AS dyad. Finally, matrix values are converted to a standardized z-score, which is calculated using the formula in **Equation 1**.

$$Z = \frac{(x - \mu)}{\sigma}$$

Equation 1. Computation of standardized z-score for CTs across ASs.

Equation 1 shows x as an average CT percentage measure, μ is the mean average CT percentage, and σ is the standard deviation mean average CT percentage. The z-score identifies how many standard deviations a data point is from the average mean. If the z-score is 0, values are close to the variable's average; a z-score of 1 indicates that CT percentage values are 1 standard deviation higher than the mean for that CT, values of 2 are 2 standard deviations from the mean, etc.

Differences across organs and CTann become visible. For example, heart, lung, kidney, and pancreas, when annotated with Azimuth (see **Fig. C1A**), show distinct bands of CTs with a z-score of ~1.5, some up to 5. We observe similar patterns for the heart, liver, lung, pancreas, skin, and small and large intestines in CellTypist (see **Fig. C1B**) and breast, heart, liver, lung, pancreas, prostate, skin, small and large intestines, spleen, thymus, urinary bladder, and ureter for popV (see **Fig. C1.C**). High-resolution versions of the heatmaps and the code to generate them are listed on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#z-scores-for-cts-per-organ-and-as.

Datasets per AS

The grouped bar graphs in **Fig. 5** show the coverage of HRApop Atlas Data across ASs and organs. Depicted is the number of datasets on the x-axis per AS on the y-axis (labeled by organ for better groupings). Overlaps and gaps in coverage between CTann tools (horizontal), across all organs and ASs (vertical) and between sex (color) become visible. The AS with the most registered HRApop Atlas Data is the outer cortex of the kidney (purl.obolibrary.org/obo/UBERON_0002189), with 75 datasets run through Azimuth. Sc-proteomics datasets are omitted. Counts for this graph are provided on the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information#dataset-counts.

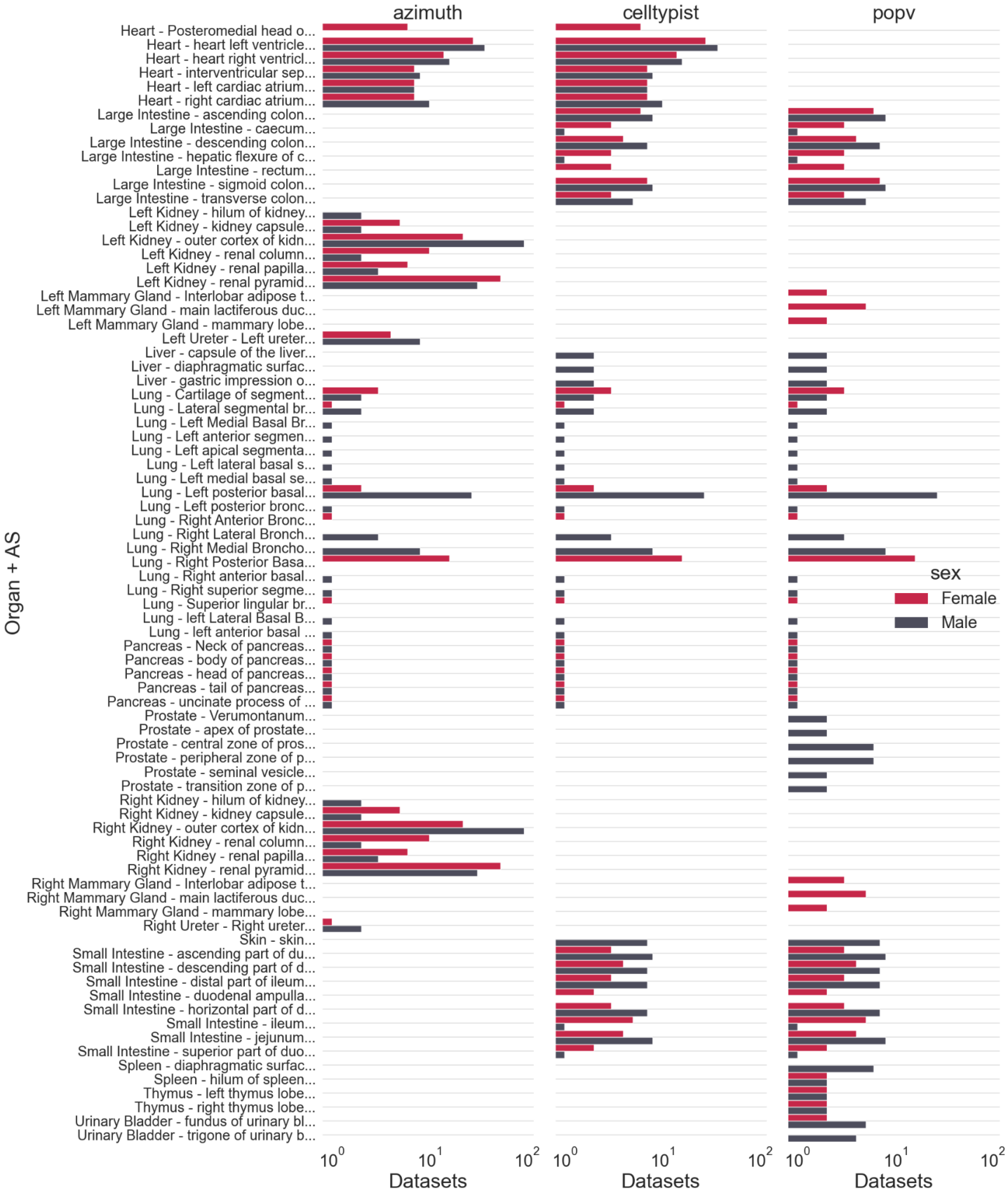

Fig. 5. HRApop Atlas Datasets per organ and AS. The number of HRApop Atlas Data per organ and AS combination (y-axis) is plotted on the x-axis, faceted by tool (columns) and colored by sex. Note that the x-axis is scaled logarithmically.

Usage Notes

Annotating local H5AD files with DCTA Workflow

While the DCTA Workflow is designed to run on H5AD files in bulk, code is available to run it on a local H5AD file provided by a user. Details are provided at cns-iu.github.io/hra-cell-type-populations-supporting-information/#running-hra-workflows-runner-on-a-local-h5ad-file. Future development for HRApop will include the ability for the user to upload H5AD files to a web server and get a CT population in return.

Getting HRApop Data via API Queries

A Jupyter Notebook detailing easy access to CT populations for ASs, extraction sites, and datasets is available at cns-iu.github.io/hra-cell-type-populations-supporting-information#accessing-hrapop-data.

Visualizing CT Populations for ASs, Extraction Sites, and Datasets

A web interface to create bar graph visualizations for HRApop Atlas CT populations is described and available at cns-iu.github.io/hra-cell-type-populations-supporting-information/#visualizing-cell-type-populations-for-as-es-and-datasets.

No-code Access to HRApop Data and Functionality

No-code UIs were implemented to allow users to enter a 3D extraction site and get a predicted CT population (Cell Population Predictor, HRA US#1) or enter a CT population and get a predicted spatial origin (Tissue Origin Predictor, HRA US#2). These predictions made with HRApop are at the AS, not the organ level. Details for both UIs are provided below.

Cell Population Predictor (US#1)

The demo application at apps.humanatlas.io/us1 uses the ASpop to predict a CT population for a user-provided extraction site with unknown CTs and numbers. A screenshot is shown in **Fig. S3.**

The user can either (a) create an extraction site with the RUI, (b) upload an existing one, or (c) use a sample extraction site[129] from the male, left kidney. After clicking the Generate Summary button, the user is presented with a CT population based on the extraction site. Behind the scenes, the application sends the extraction site to the HRA Mesh Collision API, gets the AS tags and intersection percentages (see **Box 1**), then calculates the resulting CT population based on the ASpop for any colliding AS for which HRApop data exists. As a result, the user gets a table with one row per CT, and with columns for the CTann tool, the modality, the percentage of that CT for the total summary, a raw count, a cell label, and a cell ID. The API call used by the demo is documented at apps.humanatlas.io/api/#post-/hra-pop/rui-location-cell-summary. The code for this application is available on GitHub[130]. HRA US#1 can be used to (1) predict CT populations for datasets for which no assays have been performed yet or (2) perform QA/QC on an assay by validating its results against a query in HRApop. The code and API calls used by this demo are documented in **Table S2.**

Tissue Origin Predictor (US#2)

The demo application at apps.humanatlas.io/us2 uses the ASpop and the DESpop to predict the most similar ASs for which HRApop data exist, the most similar extraction sites, and the most similar HRApop datasets for a user-provided CT population of a dataset with unknown spatial origin. A screenshot is shown in **Fig. S4**.

The user can either (a) upload a CSV file with a CT population for a dataset or (b) use a sample CSV file[131] with a CT population for the heart. Then, the user is asked to choose the organ of origin from a dropdown list, as well as the CTann tool with which the CT population was created (Azimuth, CellTypist, or popV). When clicking the Generate Predictions button, three types of possible spatial origin are presented, all sorted by weighted cosine similarity (see **Methods** section) between the AS and the user-provided CT population: (1) a list of similar ASs, (2) a list of similar datasets; and (3) a link to an EUI with the most similar extraction sites. To retrieve these, the weighted cosine similarity between the vector representing the user-provided CT population and the HRApop ASpop is computed. This is a value between 0 (no shared CT counts) and 1 (same CT counts). The API call used by the demo is documented at apps.humanatlas.io/api/#post-/hra-pop/rui-location-cell-summary. The code for this application is available on GitHub[132]. The code and API calls used by this demo are documented in **Table S2.**

Code Availability

An overview of all GitHub repositories used to construct and use HRApop is provided in **Table S1**, separated by Data, Code (Construction), Code (Support), Collision Detection and Corridors, and Coverage and Visualization. All HRA UIs that expose/use HRApop are listed in **Table S2**. A companion website for this paper is available at cns-iu.github.io/hra-cell-type-populations-supporting-information and hosted in a GitHub repository[133].

The Sankey diagram from **Fig. 3** can be explored interactively at cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_universe_plotly.html. A version showing only HRApop Atlas Data is shown in **Fig. S1**, with an interactive version available at cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_atlas_plotly.html.

References

1. Mccrory, M. A., Gomez, T. D., Bernauer, E. M. & Molé, P. A. Evaluation of a new air displacement plethysmograph for measuring human body composition. *Med. Sci. Sports Exerc.* **27**, 1686 (1995).
2. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).
3. Hatton, I. A. *et al.* The human cell count and size distribution. *Proc. Natl. Acad. Sci.* **120**, e2303077120 (2023).
4. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
5. The Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
6. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).
7. Osumi-Sutherland, D. *et al.* Cell type ontologies of the Human Cell Atlas. *Nat. Cell Biol.* **23**, 1129–1135 (2021).
8. Jain, S. *et al.* Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). *Nat. Cell*

Biol. **25**, 1089–1100 (2023).
9. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
10. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
11. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
12. Rood, J. E. *et al.* The Human Cell Atlas from a cell census to a unified foundation model. *Nature* **637**, 1065–1071 (2025).
13. Erö, C., Gewaltig, M.-O., Keller, D. & Markram, H. A Cell Atlas for the Mouse Brain. *Front. Neuroinformatics* **12**, (2018).
14. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
15. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
16. CZI Cell Science Program *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
17. Bidanta, S. *et al.* Functional tissue units in the Human Reference Atlas. *Nat. Commun.* **16**, 1526 (2025).
18. Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **46**, D246–D251 (2018).
19. Moreno, P. *et al.* Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.* **50**, D129–D140 (2022).
20. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).
21. El-Achkar, T. M. *et al.* A multimodal and integrated approach to interrogate human kidney biopsies with rigor and reproducibility: guidelines from the Kidney Precision Medicine Project. *Physiol. Genomics* **53**, 1–11 (2021).
22. Bueckle, A. *et al.* Construction, Deployment, and Usage of the Human Reference Atlas Knowledge Graph. *Sci. Data* **12**, 1100 (2025).
23. SenNet Consortium *et al.* NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. *Nat. Aging* **2**, 1090–1100 (2022).
24. Börner, K. *et al.* Tissue registration and exploration user interfaces in support of a human reference atlas. *Commun. Biol.* **5**, 1369 (2022).
25. Bueckle, A., Buehling, K., Shih, P. C. & Börner, K. 3D virtual reality vs. 2D desktop registration user interface comparison. *PLOS ONE* **16**, e0258103 (2021).
26. Bueckle, A. & Qaurooni Fard, D. Using the Standalone Registration User Interface. (2024) doi:10.5281/zenodo.14346543.
27. Bueckle, A. & Qaurooni Fard, D. Using the Embedded Registration User Interface. (2024) doi:10.5281/zenodo.14346342.
28. Qaurooni Fard, D. & Wright, D. Managing Human Reference Atlas Registrations. (2024) doi:10.5281/zenodo.11625379.
29. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
30. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
31. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
32. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**, 555–565 (2022).
33. Israel, U. *et al.* CellSAM: A Foundation Model for Cell Segmentation. 2023.11.17.567630 Preprint at https://doi.org/10.1101/2023.11.17.567630 (2025).
34. Börner, K. *et al.* Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas construction

and usage. *Nat. Methods* 1–16 (2025) doi:10.1038/s41592-024-02563-5.
35. Börner, K. *et al.* Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nat. Cell Biol.* **23**, 1117–1128 (2021).
36. National Library of Medicine. The Visible Human Project. https://www.nlm.nih.gov/research/visible/getting_data.html (2023).
37. Spitzer, V., Ackerman, M. J., Scherzinger, A. L. & Whitlock, D. The visible human male: a technical report. *J. Am. Med. Inform. Assoc.* **3**, 118–130 (1996).
38. Ackerman, M. J. The Visible Human project. *Proc. IEEE* **86**, 504–511 (1998).
39. Jain, Y. *et al.* Segmenting functional tissue units across human organs using community-driven development of generalizable machine learning algorithms. *Nat. Commun.* **14**, 4656 (2023).
40. Ma, J. *et al.* Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
41. Jain, Y. *et al.* Segmentation of human functional tissue units in support of a Human Reference Atlas. *Commun. Biol.* **6**, 717 (2023).
42. Wang, X. (Julie) *et al.* Generalized cell phenotyping for spatial proteomics with language-informed vision models. 2024.11.02.621624 Preprint at https://doi.org/10.1101/2024.11.02.621624 (2024).
43. Israel, U. *et al.* A Foundation Model for Cell Segmentation. Preprint at https://doi.org/10.48550/arXiv.2311.11004 (2023).
44. Jain, Y. *et al.* Vasculature segmentation in 3D hierarchical phase-contrast tomography images of human kidneys. 2024.08.25.609595 Preprint at https://doi.org/10.1101/2024.08.25.609595 (2024).
45. Yagis, E. *et al.* Deep learning for 3D vascular segmentation in hierarchical phase contrast tomography: a case study on kidney. *Sci. Rep.* **14**, 27258 (2024).
46. Brbić, M. *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nat. Methods* **19**, 1411–1418 (2022).
47. Yu, T. *et al.* Tissue reassembly with generative AI. 2025.02.13.638045 Preprint at https://doi.org/10.1101/2025.02.13.638045 (2025).
48. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
49. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
50. Xu, C. *et al.* Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* **186**, 5876-5891.e20 (2023).
51. Ergen, C. *et al.* Consensus prediction of cell type labels in single-cell data with popV. *Nat. Genet.* (2024) doi:10.1038/s41588-024-01993-3.
52. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **7**, 44 (2016).
53. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
54. Meehan, T. F. *et al.* Logical Development of the Cell Ontology. *BMC Bioinformatics* **12**, 6 (2011).
55. Puig-Barbe, A. hra-pop/src/utils/slim-hierarchy.csv at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/blob/main/src/utils/slim-hierarchy.csv (2025).
56. Quardokus, E. M. *et al.* Organ Mapping Antibody Panels: a community resource for standardized multiplexed tissue imaging. *Nat. Methods* **20**, 1174–1178 (2023).
57. Qaurooni, D., Herr II, B. W., Wright, D. & Bueckle, A. HRA Registrations: Manually curated HRA Dataset Graphs. https://github.com/hubmapconsortium/hra-registrations (2025).
58. Herr II, B. W. & Bueckle, A. hra-pop/output-data/v1.0/reports at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/tree/main/output-data/v1.0/reports (2025).
59. Michael Hausenblas. 5-star Open Data. http://5stardata.info/en/ (2024).
60. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
61. Cyberinfrastructure for Network Science Center. hubmapconsortium/hra-workflows-runner. https://github.com/hubmapconsortium/hra-workflows-runner (2025).

62. Cyberinfrastructure for Network Science Center. HRA workflows for processing experimental data. https://github.com/hubmapconsortium/hra-workflows (2025).
63. Tan, S. Z. K. *et al.* Brain Data Standards - A method for building data-driven cell-type ontologies. *Sci. Data* **10**, 50 (2023).
64. Cyberinfrastructure for Network Science Center. hra-pop. https://github.com/x-atlas-consortia/hra-pop (2025).
65. Bueckle, A. *et al.* The HRA Organ Gallery affords immersive superpowers for building and exploring the Human Reference Atlas with virtual reality. *Front. Bioinforma.* **3**, (2023).
66. Andreas Bueckle. HRA Organ Gallery on Meta Horizon Store. *Meta Quest Store* https://www.meta.com/en-gb/experiences/hra-organ-gallery/5696814507101529/ (2025).
67. Ardini-Poleske, M. E. *et al.* LungMAP: The Molecular Atlas of Lung Development Program. *Am. J. Physiol.-Lung Cell. Mol. Physiol.* **313**, L733–L740 (2017).
68. Gaddis, N. *et al.* LungMAP Portal Ecosystem: Systems-level Exploration of the Lung. *Am. J. Respir. Cell Mol. Biol.* **70**, 129–139 (2024).
69. Bueckle, A., Herr II, B. W. & Börner, K. HRApop v1.0. Zenodo https://doi.org/10.5281/zenodo.15603820 (2025).
70. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).
71. Lin, J.-R. *et al.* Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer. *Cell* **186**, 363-381.e19 (2023).
72. Ghose, S. *et al.* 3D reconstruction of skin and spatial mapping of immune cell density, vascular distance and effects of sun exposure and aging. *Commun. Biol.* **6**, 718 (2023).
73. dos Santos Peixoto, R. *et al.* Characterizing cell-type spatial relationships across length scales in spatially resolved omics data. *Nat. Commun.* **16**, 350 (2025).
74. Currlin, S. *et al.* Immune, endothelial and neuronal network map in human lymph node and spleen. 2021.10.20.465151 Preprint at https://doi.org/10.1101/2021.10.20.465151 (2022).
75. Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y. & Sorger, P. K. Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. *Curr. Protoc. Chem. Biol.* **8**, 251–264 (2016).
76. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci.* **110**, 11982–11987 (2013).
77. Sood, A. *et al.* Comparison of Multiplexed Immunofluorescence Imaging to Chromogenic Immunohistochemistry of Skin Biomarkers in Response to Monkeypox Virus Infection. *Viruses* **12**, 787 (2020).
78. McDonough, L., Chadwick, C., Ginty, F., Surrette, C. & Sood, A. Cell DIVE™ Platform | Antibody Characterization for Multiplexing. (2020).
79. Black, S. *et al.* CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.* **16**, 3802–3835 (2021).
80. Radtke, A. J. *et al.* IBEX: an iterative immunolabeling and chemical bleaching method for high-content imaging of diverse tissues. *Nat. Protoc.* **17**, 378–401 (2022).
81. Radtke, A. J. *et al.* IBEX: A versatile multiplex optical imaging approach for deep phenotyping and spatial analysis of cells in complex tissues. *Proc. Natl. Acad. Sci.* **117**, 33455–33465 (2020).
82. Nature Methods Editorial. Method of the Year 2024: spatial proteomics. *Nat. Methods* **21**, 2195–2196 (2024).
83. Cyberinfrastructure for Network Science Center. hra-node-dist-vis/docs/datasets.json at main · cns-iu/hra-node-dist-vis. *GitHub* https://github.com/cns-iu/hra-node-dist-vis/blob/main/docs/datasets.json (2025).
84. Cyberinfrastructure for Network Science Center. hra-pop/input-data/v1.0/sc-proteomics-cell-summaries.jsonld at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/blob/main/input-data/v1.0/sc-proteomics-cell-summaries.jsonld (2025).

85. Cyberinfrastructure for Network Science Center. hra-node-dist-vis/docs/sc-proteomics-unmapped-labels.csv at main · cns-iu/hra-node-dist-vis. *GitHub* https://github.com/cns-iu/hra-node-dist-vis/blob/main/docs/sc-proteomics-unmapped-labels.csv (2025).
86. Zilbauer, M. *et al.* A Roadmap for the Human Gut Cell Atlas. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 597–614 (2023).
87. Li, M. *et al.* DISCO: a database of Deeply Integrated human Single-Cell Omics data. *Nucleic Acids Res.* **50**, D596–D602 (2022).
88. Weber, G. M., Ju, Y. & Börner, K. Considerations for Using the Vasculature as a Coordinate System to Map All the Cells in the Human Body. *Front. Cardiovasc. Med.* **7**, (2020).
89. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).
90. Hu, J. *et al.* Benchmarking single cell transcriptome matching methods for incremental growth of reference atlases. 2025.04.10.648034 Preprint at https://doi.org/10.1101/2025.04.10.648034 (2025).
91. Zhang, Y., Aevermann, B., Gala, R. & Scheuermann, R. H. Cell type matching in single-cell RNA-sequencing data using FR-Match. *Sci. Rep.* **12**, 9996 (2022).
92. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
93. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
94. Yan, W. *et al.* Cell Atlas of The Human Fovea and Peripheral Retina. *Sci. Rep.* **10**, 9802 (2020).
95. Li, J. *et al.* Integrated multi-omics single cell atlas of the human retina. *Res. Sq.* rs.3.rs-3471275 (2023) doi:10.21203/rs.3.rs-3471275/v1.
96. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
97. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
98. Guo, M. *et al.* Guided construction of single cell reference for human and mouse lung. *Nat. Commun.* **14**, 4566 (2023).
99. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**, R5 (2012).
100. Haendel, M. A. *et al.* Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semant.* **5**, 21 (2014).
101. Baker, T. *et al.* Key choices in the design of Simple Knowledge Organization System (SKOS). *J. Web Semant.* **20**, 35–49 (2013).
102. Alistair Miles & Sean Bechhofer. SKOS Simple Knowledge Organization System Reference. https://www.w3.org/TR/skos-reference/ (2009).
103. obophenotype. cell-ontology/src/templates/general_cell_types_upper_slim.csv at master · obophenotype/cell-ontology. *GitHub* https://github.com/obophenotype/cell-ontology/blob/master/src/templates/general_cell_types_upper_slim.csv (2025).
104. Cyberinfrastructure for Network Science Center. hra-pop/output-data/v1.0/reports/atlas-ad-hoc/unmapped-cell-ids.csv at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/unmapped-cell-ids.csv (2025).
105. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/util/handler.js at main · hubmapconsortium/hra-workflows-runner. *GitHub* https://github.com/hubmapconsortium/hra-workflows-runner/blob/main/src/util/handler.js (2025).
106. Cyberinfrastructure for Network Science Center. hra-workflows/containers/azimuth at main · hubmapconsortium/hra-workflows. *GitHub* https://github.com/hubmapconsortium/hra-workflows/tree/main/containers/azimuth (2025).
107. Cyberinfrastructure for Network Science Center. hra-workflows/containers/azimuth/context/organ-

metadata.json at main · hubmapconsortium/hra-workflows. *GitHub* https://github.com/hubmapconsortium/hra-workflows/blob/main/containers/azimuth/context/organ-metadata.json (2025).
108. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
109. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
110. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).
111. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
112. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
113. Cyberinfrastructure for Network Science Center. hubmapconsortium/hra-rui-locations-processor. https://github.com/hubmapconsortium/hra-rui-locations-processor (2025).
114. Cyberinfrastructure for Network Science Center. Deployed, manually curated HRA Dataset Graphs. *hra-registrations* https://hubmapconsortium.github.io/hra-registrations/ (2025).
115. Chen, L. *et al.* Real-Time Spatial Registration for 3D Human Atlas. in *Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* 27–35 (Association for Computing Machinery, New York, NY, USA, 2022). doi:10.1145/3557917.3567618.
116. Cyberinfrastructure for Network Science Center. hra-pop/src/utils/cell-summary-similarity.js at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/blob/main/src/utils/cell-summary-similarity.js (2025).
117. Cyberinfrastructure for Network Science Center. hra-pop/input-data/v1.0 at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/tree/main/input-data/v1.0 (2025).
118. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/xconsortia/metadata.js at 4e38ef2f813287583f93a0955bd47f1fdd10b3ec · hubmapconsortium/hra-workflows-runner. *GitHub* https://github.com/hubmapconsortium/hra-workflows-runner/blob/4e38ef2f813287583f93a0955bd47f1fdd10b3ec/src/xconsortia/metadata.js (2025).
119. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/gtex/downloader.js at 4e38ef2f813287583f93a0955bd47f1fdd10b3ec · hubmapconsortium/hra-workflows-runner. *GitHub* https://github.com/hubmapconsortium/hra-workflows-runner/blob/4e38ef2f813287583f93a0955bd47f1fdd10b3ec/src/gtex/downloader.js (25AD).
120. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/combine-metadata.js at 4e38ef2f813287583f93a0955bd47f1fdd10b3ec · hubmapconsortium/hra-workflows-runner. *GitHub* https://github.com/hubmapconsortium/hra-workflows-runner/blob/4e38ef2f813287583f93a0955bd47f1fdd10b3ec/src/combine-metadata.js (2025).
121. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/sennet/metadata.js at 4e38ef2f813287583f93a0955bd47f1fdd10b3ec · hubmapconsortium/hra-workflows-runner. https://github.com/hubmapconsortium/hra-workflows-runner/blob/4e38ef2f813287583f93a0955bd47f1fdd10b3ec/src/sennet/metadata.js#L64-L71 (2025).
122. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/cellxgene/extract_donor_metadata.py at main · hubmapconsortium/hra-workflows-runner. *GitHub* https://github.com/hubmapconsortium/hra-workflows-runner/blob/main/src/cellxgene/extract_donor_metadata.py (2025).
123. Cyberinfrastructure for Network Science Center. hra-workflows-runner/src/cellxgene/extract_dataset_multi.py at main · hubmapconsortium/hra-workflows-runner. https://github.com/hubmapconsortium/hra-workflows-runner/blob/main/src/cellxgene/extract_dataset_multi.py (2025).
124. Cyberinfrastructure for Network Science Center. hra-pop/scripts at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/tree/main/scripts (2025).

125. Cyberinfrastructure for Network Science Center. hra-pop/output-data/v1.0 at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/tree/main/output-data/v1.0 (2025).
126. Portaneri, C., Rouxel-Labbé, M., Hemmer, M., Cohen-Steiner, D. & Alliez, P. Alpha wrapping with an offset. *ACM Trans Graph* **41**, (2022).
127. Cyberinfrastructure for Network Science Center. hubmapconsortium/hra-corridor-generation: hra-corridor-generation: API to generate corridors (GLB) for RUI locations. https://github.com/hubmapconsortium/hra-corridor-generation (2025).
128. Wood, J. Filter and Refine Strategy. in *Encyclopedia of GIS* (eds. Shekhar, S. & Xiong, H.) 320–320 (Springer US, Boston, MA, 2008). doi:10.1007/978-0-387-35973-1_415.
129. Cyberinfrastructure for Network Science Center. hra-apps/applications/us1-spatial-to-cell/registration-data.json at main · x-atlas-consortia/hra-apps. *GitHub* https://github.com/x-atlas-consortia/hra-apps/blob/main/applications/us1-spatial-to-cell/registration-data.json (2025).
130. Cyberinfrastructure for Network Science Center. hra-apps/applications/us1-spatial-to-cell at main · x-atlas-consortia/hra-apps. *GitHub* https://github.com/x-atlas-consortia/hra-apps/tree/main/applications/us1-spatial-to-cell (2025).
131. Cyberinfrastructure for Network Science Center. hra-apps/applications/us2-cell-to-spatial/heart-cell-summary.csv at main · x-atlas-consortia/hra-apps. *GitHub* https://github.com/x-atlas-consortia/hra-apps/blob/main/applications/us2-cell-to-spatial/heart-cell-summary.csv (2025).
132. Cyberinfrastructure for Network Science Center. hra-apps/applications/us2-cell-to-spatial at main · x-atlas-consortia/hra-apps. *GitHub* https://https://github.com/x-atlas-consortia/hra-apps/tree/main/applications/us2-cell-to-spatial (2025).
133. Cyberinfrastructure for Network Science Center. cns-iu/hra-cell-type-populations-supporting-information. https://github.com/cns-iu/hra-cell-type-populations-supporting-information (2025).
134. McMahon, A. P. *et al.* GUDMAP: The Genitourinary Developmental Molecular Anatomy Project. *J. Am. Soc. Nephrol.* **19**, 667 (2008).
135. Beckett, D., Berners-Lee, T., Prud'hommeaux, E. & Carothers, G. RDF 1.2 Turtle. *RDF 1.2 Turtle* https://www.w3.org/TR/rdf12-turtle/ (2025).
136. Cyberinfrastructure for Network Science Center. hra-workflows/containers at main · hubmapconsortium/hra-workflows. *GitHub* https://github.com/hubmapconsortium/hra-workflows/tree/main/containers (2025).
137. Cyberinfrastructure for Network Science Center. hra-pop/output-data/v1.0/reports/universe-ad-hoc/unmapped-cell-ids.csv at main · x-atlas-consortia/hra-pop. *GitHub* https://github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/universe-ad-hoc/unmapped-cell-ids.csv (2025).
138. Cyberinfrastructure for Network Science Center. HRA Organ Gallery on Horizon Store. *Oculus* https://www.meta.com/experiences/quest/5696814507101529/ (2024).

Author Contributions

A.B. led the HRApop effort and the writing of this paper, compiled the figures, and provided input on the development of the DCTA Workflow. K.B. led the HRApop prototype development, led HRApop validation studies with support by M.G. and A.B., and co-wrote the manuscript. A.B. and B.W.H. implemented the RUI2CTpop Workflow. B.W.H. led the technical development of the DCTA and RUI2CT Workflows. L.C. implemented mesh-based collision detection and corridor generation. D.B. implemented the DCTA Workflow. D.Q. worked closely with tissue providers to register tissue blocks with the RUI, ran expert interviews, and developed and wrote up use cases. Y.J. provided guidance on sc-proteomics datasets as a generalized use case for HRApop. A.P.B. maintained and updated crosswalks from CT labels from CTann tools and author-provided CT labels for sc-proteomics datasets to CL plus higher-level CT aggregations. K.A. provided expert guidance on GTEx data analysis and usage. F.W. consulted on mesh-based collision detection and corridor generation.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors would like to acknowledge work by Fauzan Isnaini, Aashay Gondalia, Vikrant Deshpande, Amber Ramesh, Nikhil Mahadevaswamy, and Mahadevan Narayanan Iyer on the HRApop prototype; Ellen M. Quardokus for authoring earlier CTann crosswalks; Vicky Daiya and Humaid Ilyas for their contributions to the DCTA Workflow and initial bar graph visualizations; Keyur Parekh and Juhi Khare for designing and implementing advanced bar graph visualizations; Devin Wright and Siddharth Apte for work on RUI registrations and extraction sites; Heidi Schlehlein and Kristen Browne for designing 3D reference objects; Ushma Patel, Tracey Lynn Theriault, Libby Maier, and Juhi Khare for figure design; and Niteesha Jangam for helping compile funding information.

The authors would also like to thank Nancy Ruschman, Humaid Ilyas, Supriya Bidanta, Beverly Peng, Trang Nyuen, Thanh Long Nguyen, Richard S. Scheuermann, and Elizabeth Ginexi for their expert feedback on the near-final version of this paper, as well as Ellen M. Quardokus, Yun (Renee) Zhang, and Matthew R. Ruffalo for their feedback on earlier versions. They would also like to acknowledge expert input on CTann tools by Gesmira Molla, Skylar Li, Chuan Xu, and Can Ergen.

The HRA is under active development by HuBMAP, SenNet, KPMP, GenitoUrinary Developmental Molecular Anatomy Project (GUDMAP)[134], and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) with expert input by the HRA Editorial Board and in close collaboration with experts from 20+ other consortia. K.B. is a co-director of and is funded by the CIFAR MacMillan Multiscale Human program.

This research has been supported by the following awards, including the NIH Common Fund through the Office of Strategic Coordination/Office of the NIH Director:

- OT2OD026671 and OT2OD033756 (A.B., B.W.H., L.C., D.B., D.Q., M.G., Y.J., A.P.B., F.W., K.B.)
- U24CA268108 (A.B., B.W.H., D.B., D.Q., M.G., Y.J., K.B.)
- OT2OD030545 (A.B., B.W.H., D.B., K.B.)
- U01DK133090 (A.B., B.W.H., D.B., D.Q., K.B.)

- U2CDK114886 (A.B., B.W.H., Y.J., K.B.)
- 1R03OD039970-01 (A.B., B.W.H., D.B., M.G.)
- U24HG012090 (K.A.)
- OT2OD026675 (A.B., L.C. as NIH JumpStart Award)
- OT2OD033759 (A.B. as NIH JumpStart Fellowship)
- U24DK135157 (B.W.H., K.B.);
- U01CA242936 (L.C., F.W.)

This research uses IU's computing infrastructure which was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplemental Figures

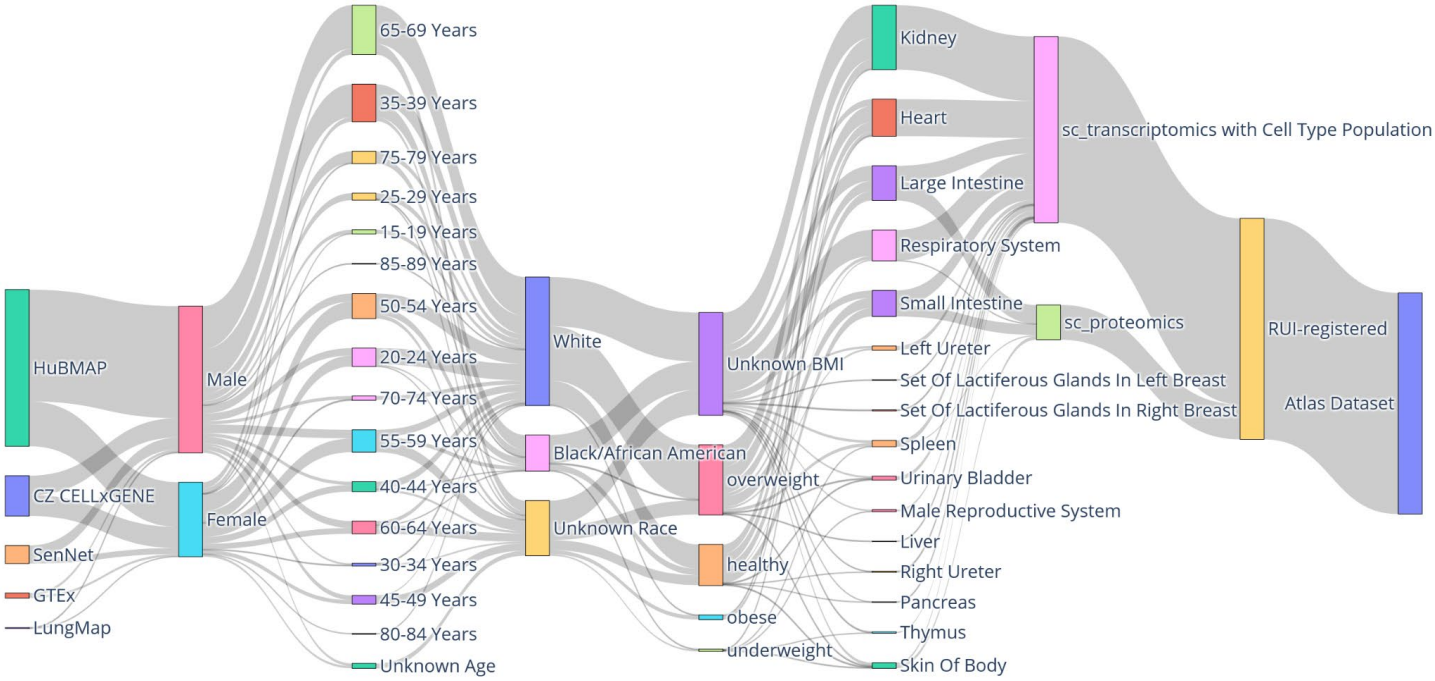

Fig. S1. A Sankey Diagram for HRApop Atlas Data. Interactive data visualization can be explored at cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_atlas_plotly.html.

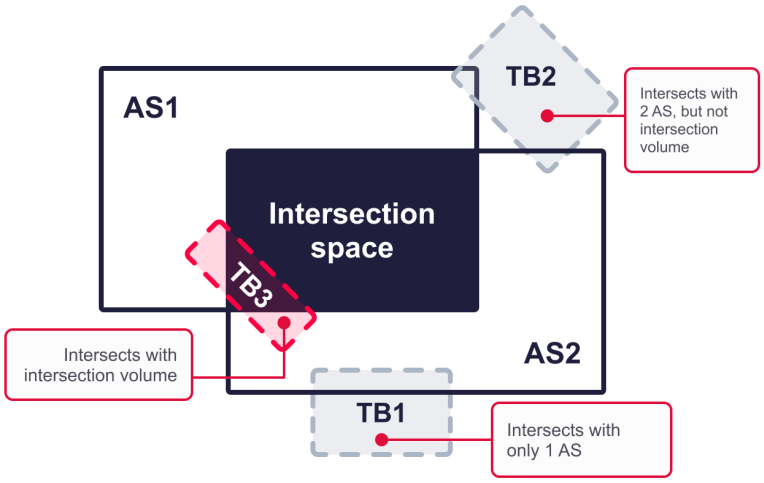

Fig. S2. **Intersection space and intersection volumes.** Two intersecting ASs create a 3D intersection space. Three exemplary tissue blocks (TB) intersect with one AS (TB1), both ASs (TB2), or both and their intersection space (TB3).

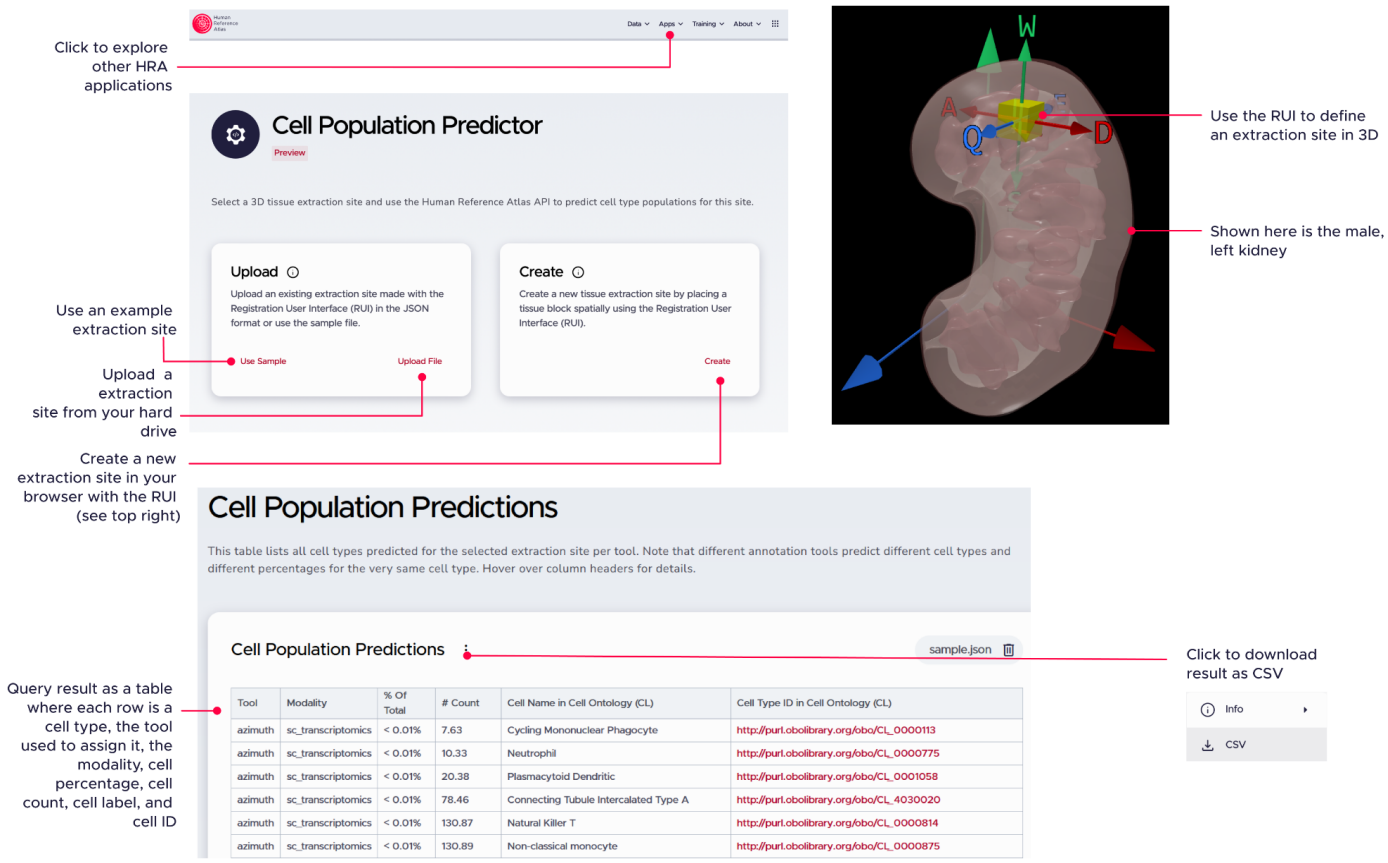

Fig. S3. Annotated screenshot of the Cell Population Predictor.

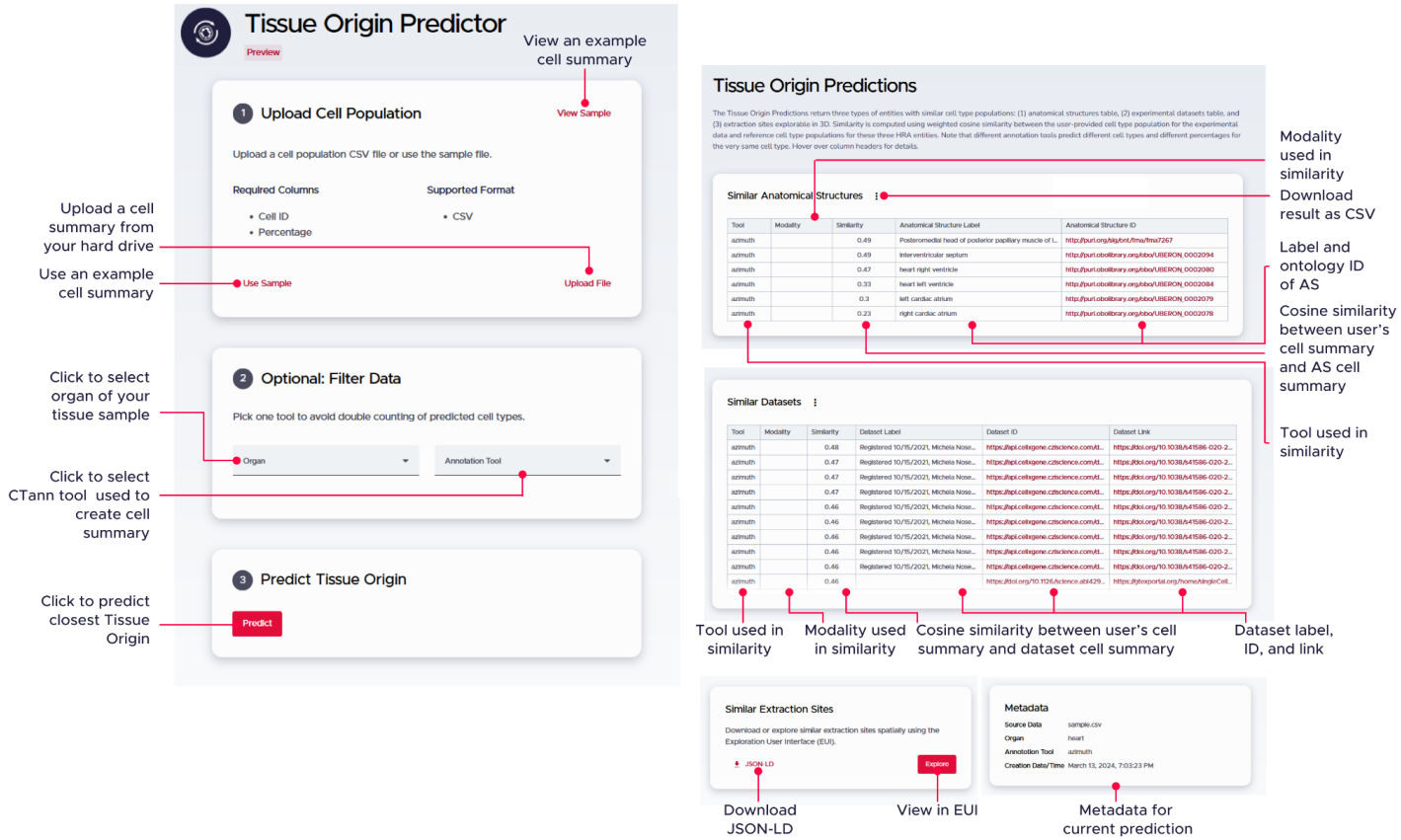

Fig. S4. Annotated screenshot of the Tissue Origin Predictor.

Supplemental Tables

Table S1. Listing of all GitHub repositories used to construct and use HRApop. Note that the HRA US#1 and US#2 Demo Applications are supported by the HRA KG, the HRA Digital Object Processor, and the HRA API. Please see details in the HRA KG paper[22].

Name	Description	URL
Data		
HRApop Atlas	ASpop and DESpop	Zenodo[69] HRA KG: purl.humanatlas.io/graph/hra-pop/v1.0
DESpop	Provided in a variety of formats (including JSON-LD and Terse RDF Triple Language [Turtle][135]) via the HRA KG and as JSON-LD on GitHub	Zenodo[69] GitHub (JSON-LD): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/atlas-enriched-dataset-graph.jsonld For datasets on GitHub (CSV): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/cell-types-per-dataset.csv For datasets on grlc.io (canned SPARQL query): apps.humanatlas.io/api/grlc/hra-pop.html#get-/cell-types-per-dataset For extraction sites on GitHub (CSV): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/cell-types-per-extraction-site.csv For extraction sites on grlc.io (canned SPARQL query): apps.humanatlas.io/api/grlc/hra-pop.html#get-/cell-types-per-extraction-site HRA KG: cdn.humanatlas.io/digital-objects/graph/hra-pop/latest/assets/atlas-enriched-dataset-graph.jsonld
ASpop	Provided via the HRA KG and via GitHub	Zenodo[69] HRA KG: cdn.humanatlas.io/digital-objects/graph/hra-pop/v1.0/assets/atlas-as-cell-summaries.jsonld GitHub (JSON-LD): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/atlas-as-cell-summaries.jsonld GitHub (CSV): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/cell-types-in-anatomical-structurescts-per-as.csv grlc.io (canned SPARQL query): apps.humanatlas.io/api/grlc/hra-pop.html#get-/cell_types_in_anatomical_structurescts_per_as

Cell instances	Cell instances for sc-transcriptomics datasets in HRApop v1.0	Zenodo[69]
Cell instances with top-10,000 biomarkers	Cell instances for sc-transcriptomics datasets in HRApop v1.0 with top-10,000 biomarkers	Zenodo[69]
Corridors	Corridors for extraction sites for HRApop v1.0	Zenodo[69] GitHub: github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/atlas-as-cell-summaries.jsonld HRA KG: lod.humanatlas.io/graph/hra-pop/v1.0/
HRApop Prediction Data	As of HRApop v1.0, there are 617 unique extraction sites for which a CT population can be predicted (US#1) and 5,453 unique datasets for which a spatial original can be predicted (US#2)	US#1: github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas/application-a1.csv US#2: github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas/application-a2p1.csv
Code (Construction)		
HRApop CTann Tool Containers	Has Docker containers for running CTann tools over H5AD files.	Repository: GitHub[62] Containers: GitHub[136]
DCTA Workflow	A set of scripts to download H5AD files, execute the Docker containers in HRApop CTann Tool Containers, and output CT populations and data metadata as input for the RUI2CTpop Workflow.	GitHub[61]
CTann crosswalks	A collection of CSV files that link CT labels from Azimuth, CellTypist, and popV to CL or PCL IDs so they can be connected to the ASCT+B tables and other HRA Digital Objects. For example, the most recent crosswalk CSV file for Azimuth is available for download at the bottom of	lod.humanatlas.io/ctann

	the HRA Digital Object landing page at lod.humanatlas.io/ctann/azimuth/latest.	
CTs level mapping	Report and queries to map crosswalked, unique CTs to higher-level CTs	Report: github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/cell-types-level-mapping.csv Query: github.com/x-atlas-consortia/hra-pop/blob/main/queries/atlas-ad-hoc/cell-types-level-mapping.rq Report (long): github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/cell-types-level-mapping-long.csv Query: github.com/x-atlas-consortia/hra-pop/blob/main/queries/atlas-ad-hoc/cell-types-level-mapping-long.rq
Not crosswalked	Reports for cell IDs that were not crosswalked during the DCTA Workflow	Input for RUI2CTpop Workflow: GitHub[137] HRApop Atlas: GitHub[104]
CT populations and metadata for sc-transcriptomics and sc-proteomics data, produced by the DCTA Workflow and serving as input for the RUI2CTpop Workflow	**sc-transcriptomics-cell-summaries.jsonld.gz**: Contains CT populations for all sc-transcriptomics datasets annotated with a CTann tool in JSON-LD. **sc-transcriptomics-dataset-metadata.csv**: Contains organ, donor metadata, . It also features a handler ID to denote which portal a dataset was downloaded from, a UUID, an assay type, and tissue provider information. **sc-proteomics-cell-summaries.jsonld:** Contains CT populations for all sc-proteomics datasets in JSON-LD. The summaries are produced via the DCTA Workflow using the code at github.com/cns-iu/hra-node-dist-vis/blob/main/scripts/build-	GitHub (all)[117] GitHub (cell summaries for sc-protemics only)[84]

	cell-summaries.js, which queries a listing of all possible sc-proteomics datasets (see GitHub[83]). **sc-transcriptomics-dataset-metadata.csv:** Contains donor, tissue block, and data IDs for sc-proteomics datasets.	
RUI2CTpop Workflow	A collection of scripts to compile HRApop from the output of the DCTA Workflow, which first provides CT populations and dataset metadata, then copies those files over to GitHub[117]. Scripts running over these input files are also on GitHub[124]. Output data is provided at github.com/x-atlas-consortia/hra-pop/tree/main/output-data/v1.0. For downstream analysis, reports based on SPARQL queries are on GitHub[58]. A full log is available at raw.githubusercontent.com/x-atlas-consortia/hra-pop/refs/heads/main/output-data/v1.0/log.txt. A file to capture the creation date of the pipeline having finished is at github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/CREATION_DATE. A Readme for the most recent run is present at github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/README.md.	GitHub[64]
Code (Support)		

HRA Registrations	Manually curated HRA Dataset Graphs. This repository holds static dataset graphs for use in the EUI and other HRA applications. All registration Digital Objects are published to hubmapconsortium.github.io/hra-registrations/. All registrations are at hubmapconsortium.github.io/hra-registrations/**name-in-root-folder**/rui_locations.jsonld	GitHub[57]
HRA Registrations Processor	Command line interface to simplify creating rui_locations.jsonld files	github.com/hubmapconsortium/hra-rui-locations-processor
Querying HRApop	A collection of SPARQL queries for use in this paper and for downstream analysis (to identify counts and get CT populations, biomarkers)	grlc.io/api-git/hubmapconsortium/ccf-grlc/subdir/hra-pop/
RUI	Stand-alone RUI use to register tissue blocks	Stand-alone RUI: apps.humanatlas.io/rui/ RUI code repository: github.com/hubmapconsortium/hra-ui/tree/main/apps/ccf-rui
EUI	Deployed stand-alone RUI extraction site sets displayed in the EUI	hubmapconsortium.github.io/hra-registrations
CTann tools	Azimuth: v0.4.6 CellTypist: v1.6 popV: github.com/YosefLab/popV/tree/2d29c9a290d2015ec65ef0ef9f0e6b6d2277e7bb	Azimuth: azimuth.hubmapconsortium.org CellTypist: www.celltypist.org popV: github.com/YosefLab/PopV
Collision Detection and Corridors		
HRA Mesh Collision API	Given an extraction site, get mesh-based collisions with the 3D reference object.	Code: github.com/hubmapconsortium/hra-tissue-block-annotation Deployed: pfn8zf2gtu.us-east-2.awsapprunner.com/get-collisions Public endpoint: apps.humanatlas.io/api/v1/collisions Documentation for endpoint: apps.humanatlas.io/api/#post-/v1/collisions Example result:

		github.com/hubmapconsortium/hra-tissue-block-annotation/blob/main/examples/test-registration-collisions.json
3D Corridor Generation API	Given an extraction site, generate a corridor with the 3D reference object as a GLB file	Code: github.com/hubmapconsortium/hra-corridor-generation Deployed: dwwcpwad72.us-east-2.awsapprunner.com/get-corridor Public endpoint: apps.humanatlas.io/api/v1/corridor Documentation for endpoint: apps.humanatlas.io/api/#post-/v1/corridor
Corridors for downloading	Contains all corridor GLB files	github.com/x-atlas-consortia/hra-pop/tree/main/output-data/v1.0/corridors
Mesh-mesh collision and annotation		Code: github.com/hubmapconsortium/hra-glb-mesh-collisions On PyPi: pypi.org/project/hra-glb-mesh-collisions/0.1.0/
Coverage and Visualization		
CT Populations by AS	Provides a query for an overview of all AS-CT combinations in ASpop of HRApop v1.0, with sex, tool, CT, and cell percentage	apps.humanatlas.io/api/grlc/hra-pop.html#get-/cell_types_in_anatomical_structurescts_per_as.
Supporting Information for this paper	Contains assets for the companion website at cns-iu.github.io/hra-cell-type-populations-supporting-information/	github.com/cns-iu/hra-cell-type-populations-supporting-information
Sankey Diagram for HRApop Atlas	Contains the deployed Sankey diagram for HRApop Atlas	cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_atlas_plotly.html
Sankey Diagram for input for RUI2CTpop Workflow	Contains the deployed Sankey diagram for the input for RUI2CTpop Workflow	cns-iu.github.io/hra-cell-type-populations-supporting-information/sankey_universe_plotly.html
HRApop Counts	Contains numbers reported in this paper	github.com/cns-iu/hra-cell-type-populations-supporting-information/blob/main/counts/hra_pop_counts.ipynb

Table S2. Listing of all HRA applications that use HRApop data.

Name	Description	URL[17]
FTU Explorer	Web-deployed UI to view and explore cell type populations from experimental datasets for 2D FTU illustrations	apps.humanatlas.io/ftu-explorer/#/
HRA Organ Gallery[65,138]	Enables an immersive view of the HRA by showing 71 reference organs and 1,100+ tissue blocks alongside CT populations in VR[65,138].	humanatlas.io/hra-organ-gallery
HRA API	Provides programmatic access to the HRA, enabling integration with other applications and services. A SPARQL endpoint for the HRA API allows users to write their own SPARQL queries. Canned queries for HRApop are available at apps.humanatlas.io/api/grlc/hra-pop.html.	API documentation: apps.humanatlas.io/api/ The HRA API on the HRA Portal: humanatlas.io/api
Cell Population Predictor	A web application where the user can (1) generate a new extraction site, (2) upload an existing extraction site in JSON-LD, or (3) use a pre-made extraction site, then receive a CT population based on HRApop data given the AS mesh collisions of the extraction site.	Deployed at: apps.humanatlas.io/us1/ On HRA Portal: humanatlas.io/user-story/1 Code: github.com/x-atlas-consortia/hra-apps/tree/main/applications/us1-spatial-to-cell HRA API endpoints used: apps.humanatlas.io/api/hra-pop/rui-location-cell-summary
Tissue Origin Predictor	A web application where the user can (1) upload their own CT population or (2) use a provided one, indicate organ and CTann tool used, and then receive (a) a list of similar ASs with CT populations (weighted cosine similarity, see **Methods** section), (b) a list of most similar datasets (weighted cosine similarity), and (c) an EUI with the most similar extraction sites.	Deployed at: apps.humanatlas.io/us2/ On HRA Portal: humanatlas.io/user-story/2 Code: github.com/x-atlas-consortia/hra-apps/tree/main/applications/us2-cell-to-spatial HRA API endpoints used: apps.humanatlas.io/api/#post-/hra-pop/cell-summary-report apps.humanatlas.io/api/#get-/hra-pop/supported-organs apps.humanatlas.io/api/#get-/hra-pop/supported-tools

Table S3. Reports about dataset and cell counts from the RUI2CTpop Workflow.

Counts		
Input for RUI2CTpop Workflow: sc-transcriptomics cell counts	Contains the number of annotated cells in sc-transcriptomics datasets that are part of the input for RUI2CTpop Workflow	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/universe-ad-hoc/universe-sc-transcriptomics-cell-counts.csv
Input for RUI2CTpop Workflow: sc-transcriptomics cell counts (with pre-annotated cell counts)	Contains the number of annotated cells in sc-transcriptomics datasets that are part of the input for RUI2CTpop Workflow that have been mapped via CTann tools (see **Box 1**). Pre-annotated cell counts are the number of rows in the cell by gene matrix of the H5AD file. Some cells may not get annotated for various reasons, thus the discrepancy	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/universe-ad-hoc/universe-sc-transcriptomics-cell-instance-counts.csv
Input for RUI2CTpop Workflow: sc-proteomics cell counts	Contains the number of annotated cells in sc-proteomics datasets that are part of the HRApop Atlas	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/universe-ad-hoc/universe-sc-proteomics-cell-counts.csv
HRApop Atlas: sc-transcriptomics cell counts	Contains the number of annotated cells in sc-transcriptomics datasets that are part of the HRApop Atlas	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/atlas-sc-transcriptomics-cell-counts.csv
HRApop Atlas: sc-proteomics cell counts	Contains the number of annotated cells in sc-proteomics datasets that are part of the HRApop Atlas	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas-ad-hoc/atlas-sc-proteomics-cell-counts.csv
HRApop Atlas: Number of extraction sites per AS and intersection volume/percentage of the extraction site (see **Box 1**)	Lists extraction site, organ/AS, intersection volume, AS volume and tissue block volume.	github.com/x-atlas-consortia/hra-pop/blob/main/output-data/v1.0/reports/atlas/table-s5.csv

SPARQL queries	Contains SPARQL queries for all reports generated in in the output data folder on GitHub[125]	github.com/x-atlas-consortia/hra-pop/tree/main/queries

Table S4. CT choices. 26 of the 201 CTs were classified under two different high-level CTs. This table shows the decision process used to select the most appropriate grouping class for each CT.

CT	Chosen Category	Non-chosen category	Rationale
hematopoietic stem cell	hematopoietic cell	stem cell	Reflects lineage identity central to its biological role
neuroendocrine cell	epithelial cell	neural cell	Developmentally and structurally epithelial
pancreatic D cell	epithelial cell	neural cell	Developmentally and structurally epithelial
intestinal crypt stem cell of large intestine	stem cell	epithelial cell	Stemness prioritized as foundational to epithelial renewal
intestinal crypt stem cell of small intestine	stem cell	epithelial cell	Stemness prioritized as foundational to epithelial renewal
lung interstitial macrophage	hematopoietic cell	connective tissue cell	Lineage from hematopoietic system prioritized over location
pulmonary neuroendocrine cell	epithelial cell	neural cell	Developmentally and structurally epithelial
granulocyte monocyte progenitor cell	hematopoietic cell	bone cell	Reflects hematopoietic lineage over location in the bone marrow
mesenchymal stem cell of adipose tissue	stem cell	connective tissue cell	Stemness and multipotency prioritized over local connective tissue association.
mesenchymal stem cell of abdominal adipose tissue	stem cell	connective tissue cell	Stemness and multipotency prioritized over local connective tissue association.
central nervous system macrophage	hematopoietic cell	neural cell	Immune lineage prioritized over CNS location
vascular leptomeningeal cell	connective tissue cell	neural cell	Mesenchymal origin and structural connective function prioritized over CNS location
mature microglial cell	hematopoietic cell	neural cell	Immune lineage prioritized over CNS location
ependymal cell	neural cell	epithelial cell	Derived from neuroectoderm and part of CNS despite epithelial morphology
smooth muscle cell of the brain vasculature	muscular cell	neural cell	Reflects muscular identity, over location
mesenchymal stem cell	stem cell	connective tissue cell	Stemness and multipotency prioritized over local connective tissue association.
basal cell of epidermis	stem cell	epithelial cell	Stemness emphasized as central to epidermal renewal

choroid plexus epithelial cell	neural cell	epithelial cell	Derived from neuroectoderm and part of CNS despite epithelial morphology
microglial cell	hematopoietic cell	neural cell	Primitive macrophage origin prioritized over CNS-resident role
type D cell of colon	epithelial cell	neural cell	Developmentally and structurally epithelial
type EC enteroendocrine cell	epithelial cell	neural cell	Developmentally and structurally epithelial
cycling type EC enteroendocrine cell	epithelial cell	neural cell	Developmentally and structurally epithelial
type D cell of small intestine	epithelial cell	neural cell	Developmentally and structurally epithelial
retinal blood vessel endothelial cell	endothelial cell	neural cell	Endothelial identity and vascular role prioritized over retinal location
retinal pigment epithelial cell	epithelial cell	neural cell	Morphology and function as barrier epithelium prioritized over retinal location
endosteal cell	bone cell	epithelial cell	Skeletal lineage and role in bone homeostasis prioritized over lining morphology

Table S5. CTann settings. Docker containers with full contexts for the three CTann tools are available on GitHub[136].

Tool	Version	Code base	Models	Requirements
Azimuth	v0.4.6	R	kidneyref/Kidney_L3/annotation.l3 lungref/Lung_v2_finest_level/ann_finest_level heartref/Heart_L2/celltype.l2 humancortexref/subclass pancreasref/Pancreas_L1/annotation.l1 pbmcref/Human_PBMC_L2/celltype.l2 bonemarrowref/Bone_marrow_L2/celltype.l2 adiposeref/Adipose_L2/celltype.l2	anndata==0.9.1 contourpy==1.0.7 cycler==0.11.0 fonttools==4.39.4 h5py==3.8.0 importlib-metadata==6.6.0 importlib-resources==5.12.0 joblib==1.2.0 kiwisolver==1.4.4 llvmlite==0.40.1rc1 matplotlib==3.7.1 natsort==8.3.1 networkx==3.1 numba==0.57.0 numpy==1.24.3 packaging==23.1 pandas==2.0.2 patsy==0.5.3 Pillow==9.5.0 pynndescent==0.5.10 pyparsing==3.0.9 python-dateutil==2.8.2 pytz==2023.3 scanpy==1.9.3 scikit-learn==1.2.2 scipy==1.10.1 seaborn==0.12.2 session-info==1.0.0 six==1.16.0 statsmodels==0.14.0 stdlib-list==0.8.0 threadpoolctl==3.1.0 tqdm==4.65.0 tzdata==2023.3 umap-learn==0.5.3 zipp==3.15.0
CellTypist	v1.6	Python	Human_Lung_Atlas.pkl Alternative: Cells_Lung_Airway.pkl Adult_Human_Skin.pkl Adult_Human_PancreaticIslet.pkl Adult_Human_PancreaticIslet.pkl Healthy_Adult_Heart.pkl Healthy_Human_Liver.pkl Human_AdultAged_Hippocampus.pkl Human_Longitudinal_Hippocampus.pkl Cells_Intestinal_Tract.pkl	anndata==0.9.2 celltypist==1.6.1 certifi==2023.7.22 charset-normalizer==3.3.1 click==8.1.7 contourpy==1.1.1 cycler==0.12.1 et-xmlfile==1.1.0 fonttools==4.43.1 h5py==3.10.0 idna==3.4 igraph==0.10.8 joblib==1.3.2 kiwisolver==1.4.5 leidenalg==0.10.1 llvmlite==0.41.1 matplotlib==3.8.0 natsort==8.4.0 networkx==3.2

				numba==0.58.1 numpy==1.24.4 openpyxl==3.1.2 packaging==23.2 pandas==2.0.3 patsy==0.5.3 Pillow==10.1.0 pynndescent==0.5.10 pyparsing==3.1.1 python-dateutil==2.8.2 pytz==2023.3.post1 requests==2.31.0 scanpy==1.9.5 scikit-learn==1.3.2 scipy==1.11.3 seaborn==0.13.0 session-info==1.0.0 six==1.16.0 statsmodels==0.14.0 stdlib-list==0.9.0 texttable==1.7.0 threadpoolctl==3.2.0 tqdm==4.66.1 tzdata==2023.3 umap-learn==0.5.4 urllib3==2.0.7
popV	See commit github.com/YosefLab/popV/tree/2d29c9a290d2015ec65ef0ef9f0e6b6d2277e7bb	Python	model: Bladder organ_level:urinary bladder model: Blood organ_level: blood model: Bone_Marrow organ_level: bone marrow model: Eye organ_level: eye model: Fat organ_level: adipose tissue model: Large_Intestine organ_level: large intestine model: Liver organ_level: liver model: Lung organ_level: lung model: Lymph_Node organ_level: mesenteric lymph node model: Mammary organ_level: mammary gland model: Pancreas organ_level: pancreas	absl-py==2.1.0 aiohappyeyeballs==2.4.0 aiohttp==3.10.5 aiosignal==1.3.1 anndata==0.10.9 annoy==1.17.3 array_api_compat==1.8 astunparse==1.6.3 attrs==24.2.0 bbknn==1.6.0 beautifulsoup4==4.12.3 celltypist==1.6.3 certifi==2024.8.30 charset-normalizer==3.3.2 chex==0.1.86 click==8.1.7 contextlib2==21.6.0 contourpy==1.3.0 cycler==0.12.1 Cython==3.0.11 docrep==0.3.2 et-xmlfile==1.1.0 etils==1.9.4 fbpca==1.0 filelock==3.15.4 flatbuffers==24.3.25 flax==0.9.0 fonttools==4.53.1 frozenlist==1.4.1 fsspec==2024.9.0 gast==0.6.0 gdown==5.2.0 geosketch==1.2 google-pasta==0.2.0

			model: Prostate Organ_level: prostate gland model: Salivary Gland model: Skin Organ_level: skin model: Small_Intestine organ_level: small intestine model: Spleen organ_level: spleen model: Thymus organ_level: thymus model: Tongue model: Trachea organ_level: trachea model: Uterus organ_level: uterus model: Vasculature organ_level: blood vasculature	grpcio==1.66.1 h5py==3.11.0 harmony-pytorch==0.1.8 huggingface-hub==0.24.6 humanize==4.10.0 idna==3.8 igraph==0.11.6 importlib_resources==6.4.4 intervaltree==3.1.0 jax==0.4.31 jaxlib==0.4.31 Jinja2==3.1.4 joblib==1.4.2 keras==3.5.0 kiwisolver==1.4.7 legacy-api-wrap==1.4 leidenalg==0.10.2 libclang==18.1.1 lightning==2.1.4 lightning-utilities==0.11.7 llvmlite==0.43.0 Markdown==3.7 markdown-it-py==3.0.0 MarkupSafe==2.1.5 matplotlib==3.9.2 mdurl==0.1.2 ml-dtypes==0.4.0 ml_collections==0.1.1 mpmath==1.3.0 msgpack==1.0.8 mudata==0.3.1 multidict==6.0.5 multipledispatch==1.0.0 namex==0.0.8 natsort==8.4.0 nest-asyncio==1.6.0 networkx==3.3 numba==0.60.0 numpy==1.26.4 numpyro==0.15.2 nvidia-cublas-cu12==12.1.3.1 nvidia-cuda-cupti-cu12==12.1.105 nvidia-cuda-nvrtc-cu12==12.1.105 nvidia-cuda-runtime- cu12==12.1.105 nvidia-cudnn-cu12==9.1.0.70 nvidia-cufft-cu12==11.0.2.54 nvidia-curand-cu12==10.3.2.106 nvidia-cusolver-cu12==11.4.5.107 nvidia-cusparse-cu12==12.1.0.106 nvidia-nccl-cu12==2.20.5 nvidia-nvjitlink-cu12==12.6.68 nvidia-nvtx-cu12==12.1.105 obonet==1.1.0 OnClass==1.3 openpyxl==3.1.5 opt-einsum==3.3.0 optax==0.2.3 optree==0.12.1 orbax-checkpoint==0.6.1 packaging==24.1

				pandas==1.5.3 patsy==0.5.6 pillow==10.4.0 PopV @ git+github.com/czbiohub/PopV@2d29c9a290d2015ec65ef0ef9f0e6b6d2277e7bb protobuf==4.25.4 psutil==6.0.0 Pygments==2.18.0 pynndescent==0.5.13 pyparsing==3.1.4 pyro-api==0.1.2 pyro-ppl==1.9.1 PySocks==1.7.1 python-dateutil==2.9.0.post0 pytorch-lightning==2.4.0 pytz==2024.1 PyYAML==6.0.2 regex==2024.7.24 requests==2.32.3 rich==13.8.0 safetensors==0.4.5 scanorama==1.7.4 scanpy==1.10.2 scikit-learn==1.1.3 scikit-misc==0.5.1 scipy==1.14.1 scvi-tools==1.1.6 seaborn==0.13.2 sentence-transformers==3.0.1 session_info==1.0.0 six==1.16.0 sortedcontainers==2.4.0 soupsieve==2.6 statsmodels==0.14.2 stdlib-list==0.10.0 sympy==1.13.2 tensorboard==2.17.1 tensorboard-data-server==0.7.2 tensorflow==2.17.0 tensorflow-io-gcs-filesystem==0.37.1 tensorstore==0.1.65 termcolor==2.4.0 texttable==1.7.0 threadpoolctl==3.5.0 tokenizers==0.19.1 toolz==0.12.1 torch==2.4.1 torchmetrics==1.4.1 tqdm==4.66.5 transformers==4.44.2 triton==3.0.0 typing_extensions==4.12.2 umap-learn==0.5.6 urllib3==2.2.2 Werkzeug==3.0.4 wget==3.2 wrapt==1.16.0 yarl==1.9.11

				zenodo-get==1.5.1 zipp==3.20.1