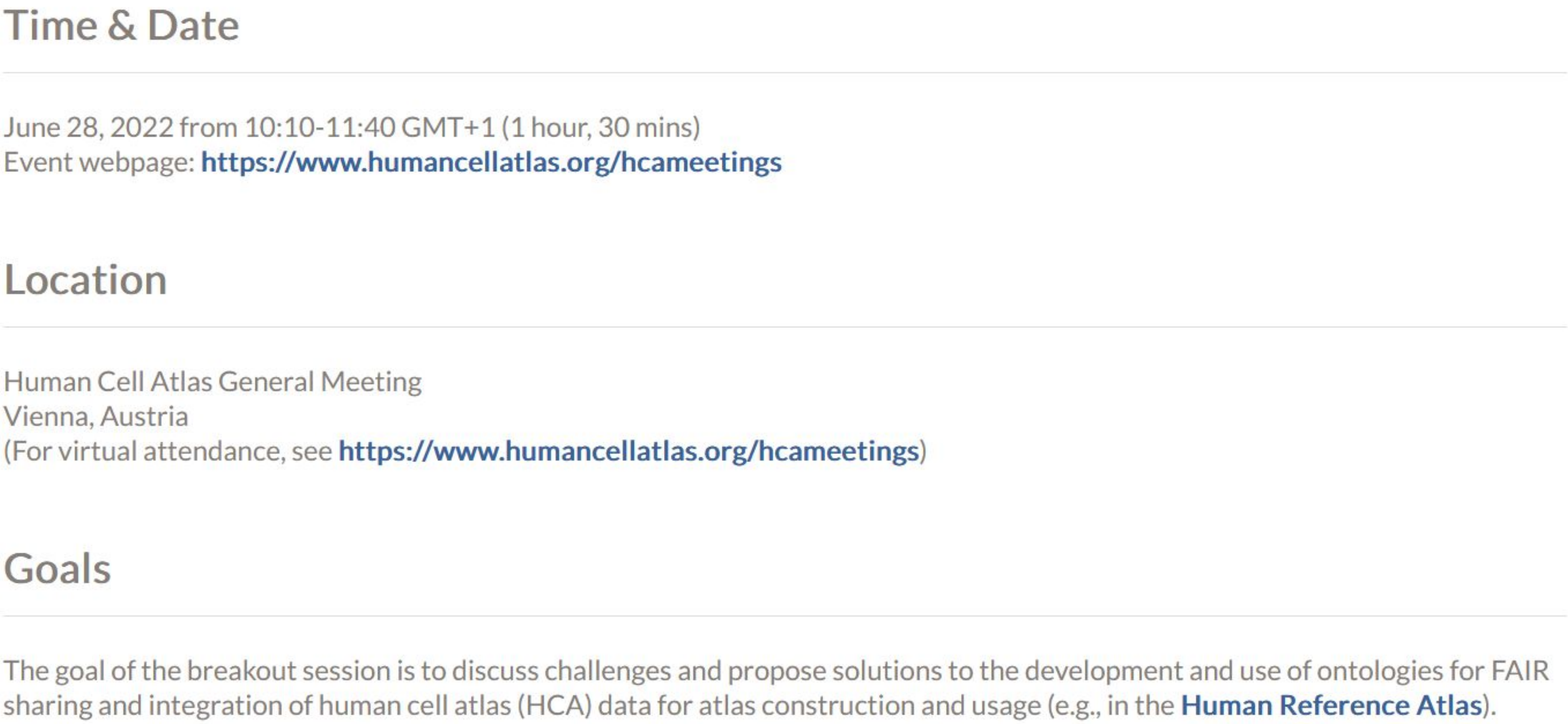
Cell Ontologies, Annotation & Metadata Session @ HCA on June 28, 2022

https://cns-iu.github.io/workshops/2022-06-27_human_cell_atlas/

https://bit.ly/HCA_Cell_ont

Overview



Time & Date

June 28, 2022 from 10:10-11:40 GMT+1 (1 hour, 30 mins)
Event webpage: **https://www.humancellatlas.org/hcameetings**

Location

Human Cell Atlas General Meeting
Vienna, Austria
(For virtual attendance, see **https://www.humancellatlas.org/hcameetings**)

Goals

The goal of the breakout session is to discuss challenges and propose solutions to the development and use of ontologies for FAIR sharing and integration of human cell atlas (HCA) data for atlas construction and usage (e.g., in the **Human Reference Atlas**).

https://cns-iu.github.io/workshops/2022-06-27_human_cell_atlas/

Organizers

Katy Börner

Indiana University, USA

katy@indiana.edu

David Osumi-Sutherland

EBI, UK

davidos@ebi.ac.uk

Key Speakers

Evan Biederstedt

Harvard Medical School, USA

Melissa Clarkson

University of Kentucky College of Medicine, USA

Bruce W. Herr II

Indiana University, USA

Slides | Video

Jason Hilton

Stanford University, USA

Wei Kheng Teh

Archival Infrastructure and Technology, EBI, UK

Angela Pisco

CZI BioHub, USA

Fabian Theis

Helmholtz-Muenchen, Germany

Summary

Standardizing the way we annotate samples and analysis metadata but also anatomical structures, cell types, and biomarkers is a key component of making HCA data Findable, Accessible, Interoperable and Re-usable (FAIR) and ultimately to integrating it into coherent atlases such as the **Human Reference Atlas (HRA)**. Ontologies, combined with standard annotation schemas, aid this process by providing standard terms for annotation and mechanisms for grouping terms in biologically meaningful ways, for example, grouping cell types by location or function. This session will discuss key challenges we face in achieving these aims and the opportunities that will be opened up by achieving them:

1. How can we **extend and improve ontologies** as our knowledge grows leveraging expert input, experimental data and feedback from different atlasing efforts?
2. How can we **make ontology annotation easy**, efficient, and accurate while leaving room for revising and adding to existing ontologies?
3. How can we enable downstream users to take advantage of ontology structure and content in **analysis, visualization and machine learning pipelines/applications**?
4. How can improved annotation with ontologies and the use of linked open data (LOD) help us to interlink atlas data and from multiple consortia and **construct more integrated, coherent, and queryable atlases**?

Agenda

Welcome and Introduction | June 28, 2022: 10.10 – 10.15 (GMT+1)

- Introduction of Workshop Goals by Session Organizers

Flash Talks | June 28, 2022: 10.15 – 10.45 (GMT+1)

- Six 5 minute flash talks on one of the challenges/opportunities, with participants talking about how their work can help &/or challenge they need, help solving

Breakout Introduction | June 28, 2022: 10.45 – 11.00 (GMT+1)

- Each group must nominate a scribe and a chair. Breakout groups will fill out structured forms for use in report back

Breakouts | June 28, 2022: 11.00 – 11.40 (GMT+1)

- Four breakouts, one on each of the four challenges/opportunities. Self assorting. Each group must nominate a scribe and a chair. Breakout groups will fill out structured form for use in report-back in the main session

Relevant Papers

- Börner, Katy, Sarah A Teichmann, Ellen M Quardokus, et al. 2021.**"Anatomical structures, cell types and biomarkers of the Human Reference Atlas"** . *Nature Cell Biology* 23: 1117-1128. doi: 10.1038/s41556-021-00788-6.
- Börner, Katy, Andreas Bueckle, Bruce W. Herr II, et al. 2021. **"Tissue Registration and Exploration User Interfaces in support of a Human Reference Atlas"**. *bioRxiv* doi: 10.1101/2021.12.30.474265.
- Osumi-Sutherland, David, Chuan Xu, Maria Keays, Adam P. Levine, Peter V. Kharchenko, Aviv Regev, Ed Lein, and Sarah A. Teichmann. 2021. "Cell Type Ontologies of the Human Cell Atlas." Nature Cell Biology 23 (11): 1129–35.
- M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci Data, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18

Please share others via https://asct-b.slack.com

Welcome

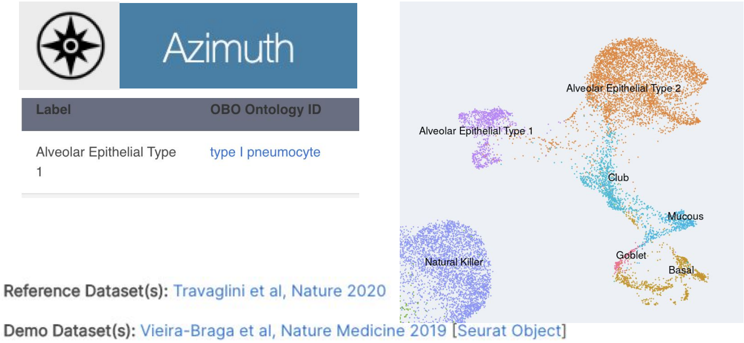
Making cell type data findable with ontology annotation

Findable Accessible Interoperable Reusable

COVID-19 Cell Atlas

wellcome sanger institute HUMAN CELL ATLAS CHAN ZUCKERBERG INITIATIVE

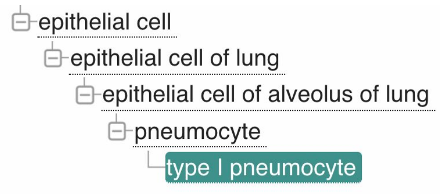
Source	Term used in data annotation
Madisoon et at., 2019 PMID:31892341	Alveolar_Type1
Lukasen et al ., 2020 DOI:10.15252/embj.20105114	AT1
Vieira Braga et al., 2019 PMID:31209336	Type_1_alveolar
Travaglini *et al.* 2020 DOI:10.1038/s41586-020-2922-4	Alveolar Epithelial Type 1



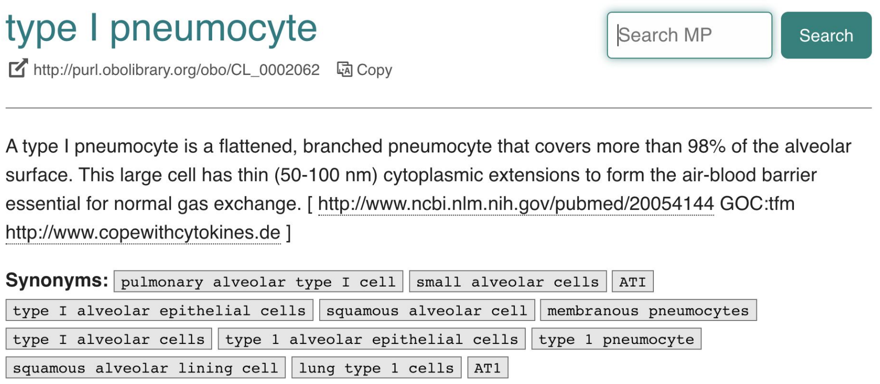
Label	OBO Ontology ID
Alveolar Epithelial Type 1	type I pneumocyte

Reference Dataset(s): Travaglini et al, Nature 2020

Demo Dataset(s): Vieira-Braga et al, Nature Medicine 2019 [Seurat Object]



epithelial cell
epithelial cell of lung
epithelial cell of alveolus of lung
pneumocyte
type I pneumocyte

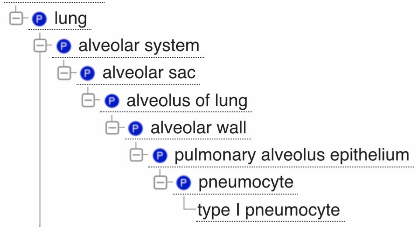


type I pneumocyte

http://purl.obolibrary.org/obo/CL_0002062 Copy

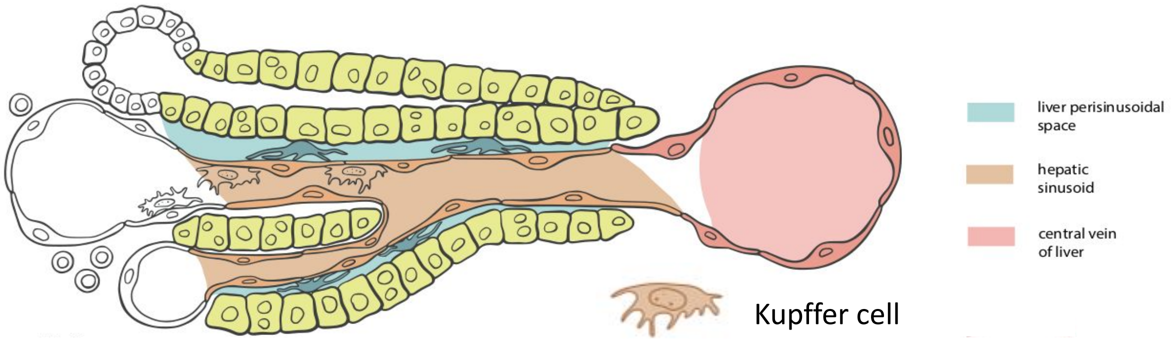
A type I pneumocyte is a flattened, branched pneumocyte that covers more than 98% of the alveolar surface. This large cell has thin (50-100 nm) cytoplasmic extensions to form the air-blood barrier essential for normal gas exchange. [http://www.ncbi.nlm.nih.gov/pubmed/20054144 GOC:tfm http://www.copewithcytokines.de]

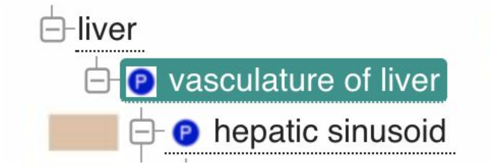
Synonyms: `pulmonary alveolar type I cell` `small alveolar cells` `ATI` `type I alveolar epithelial cells` `squamous alveolar cell` `membranous pneumocytes` `type I alveolar cells` `type 1 alveolar epithelial cells` `type 1 pneumocyte` `squamous alveolar lining cell` `lung type 1 cells` `AT1`

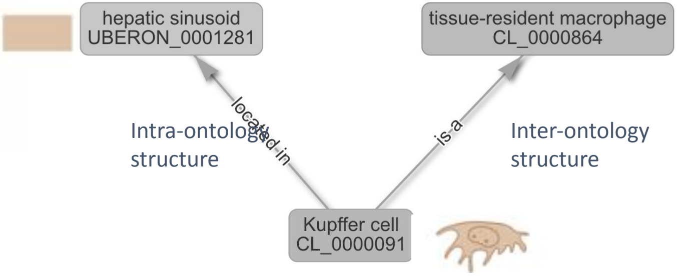


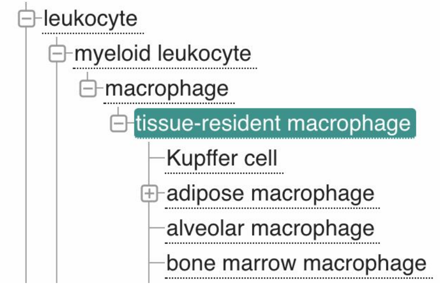
lung
alveolar system
alveolar sac
alveolus of lung
alveolar wall
pulmonary alveolus epithelium
pneumocyte
type I pneumocyte

Ontology structure allows grouping of related content – making it more *Findable* and *Interoperable*



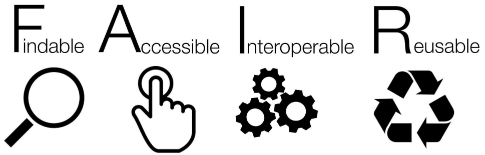

- liver
  - vasculature of liver
    - hepatic sinusoid




- leukocyte
  - myeloid leukocyte
    - macrophage
      - tissue-resident macrophage
        - Kupffer cell
        - adipose macrophage
        - alveolar macrophage
        - bone marrow macrophage

Queries:

- Find all scRNAseq datasets with data on:
  - tissue resident macrophages
  - cells in the liver



Discussion Topics

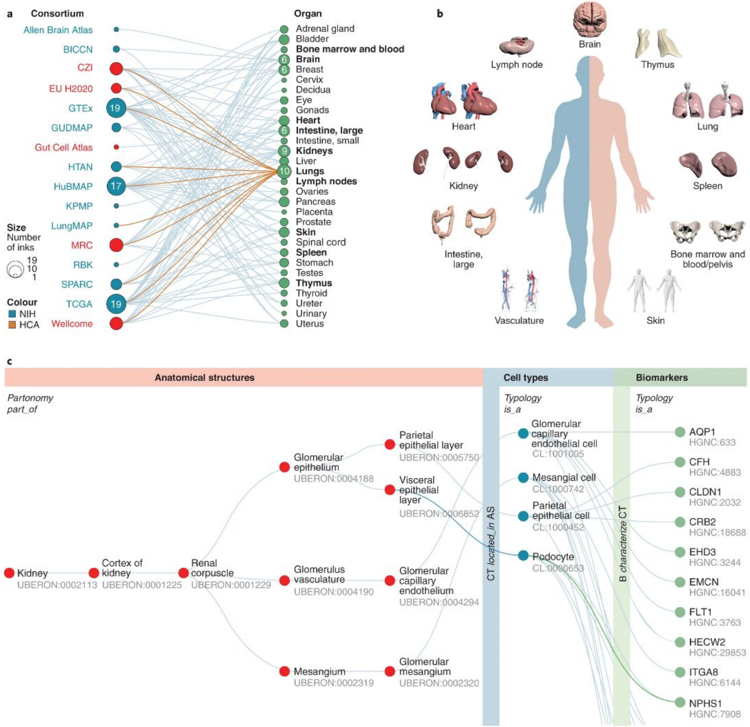
1. How can we **extend and improve ontologies** as our knowledge grows leveraging expert input, experimental data and feedback from different atlasing efforts?

   We're hiring ontology editors https://www.embl.org/jobs/position/EBI02008

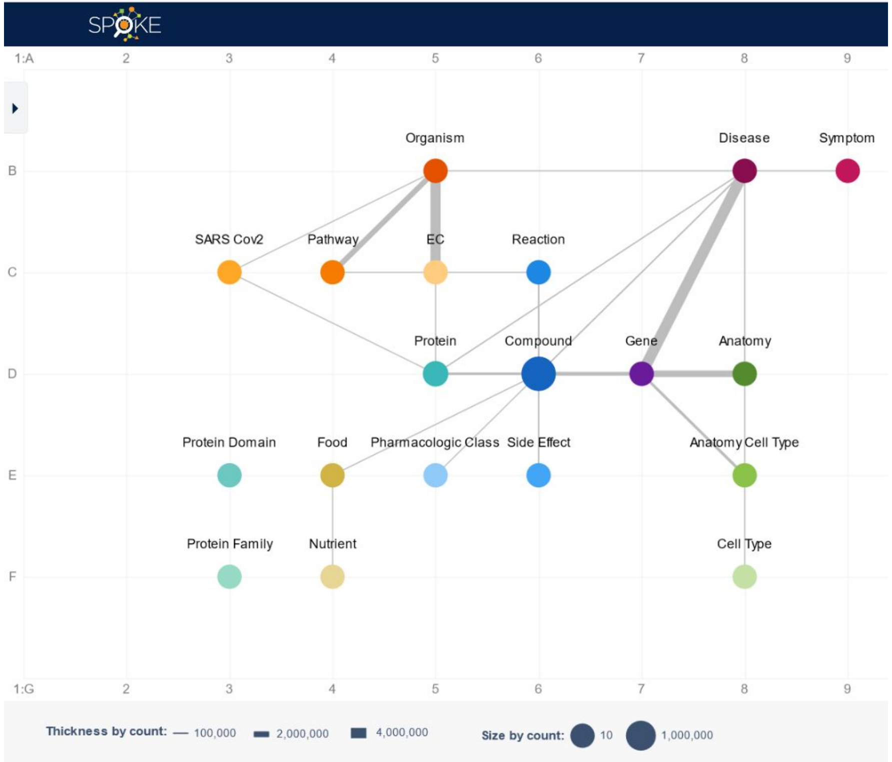
2. How can we **make ontology annotation, following standard schemas easy**, efficient, and accurate while leaving room for revising and adding to existing ontologies?
3. How can we enable downstream users to take advantage of ontology structure and content in **analysis, visualization and machine learning pipelines/applications**?
4. How can improved annotation with ontologies and the use of linked open data (LOD) help us to interlink atlas data and from multiple consortia and **construct more integrated, coherent, and queryable atlases**?

REMINDER FOR SPEAKERS – 5’ ONLY - EMPHASISE A PROBLEM/SOLUTION

Katy’s Welcome



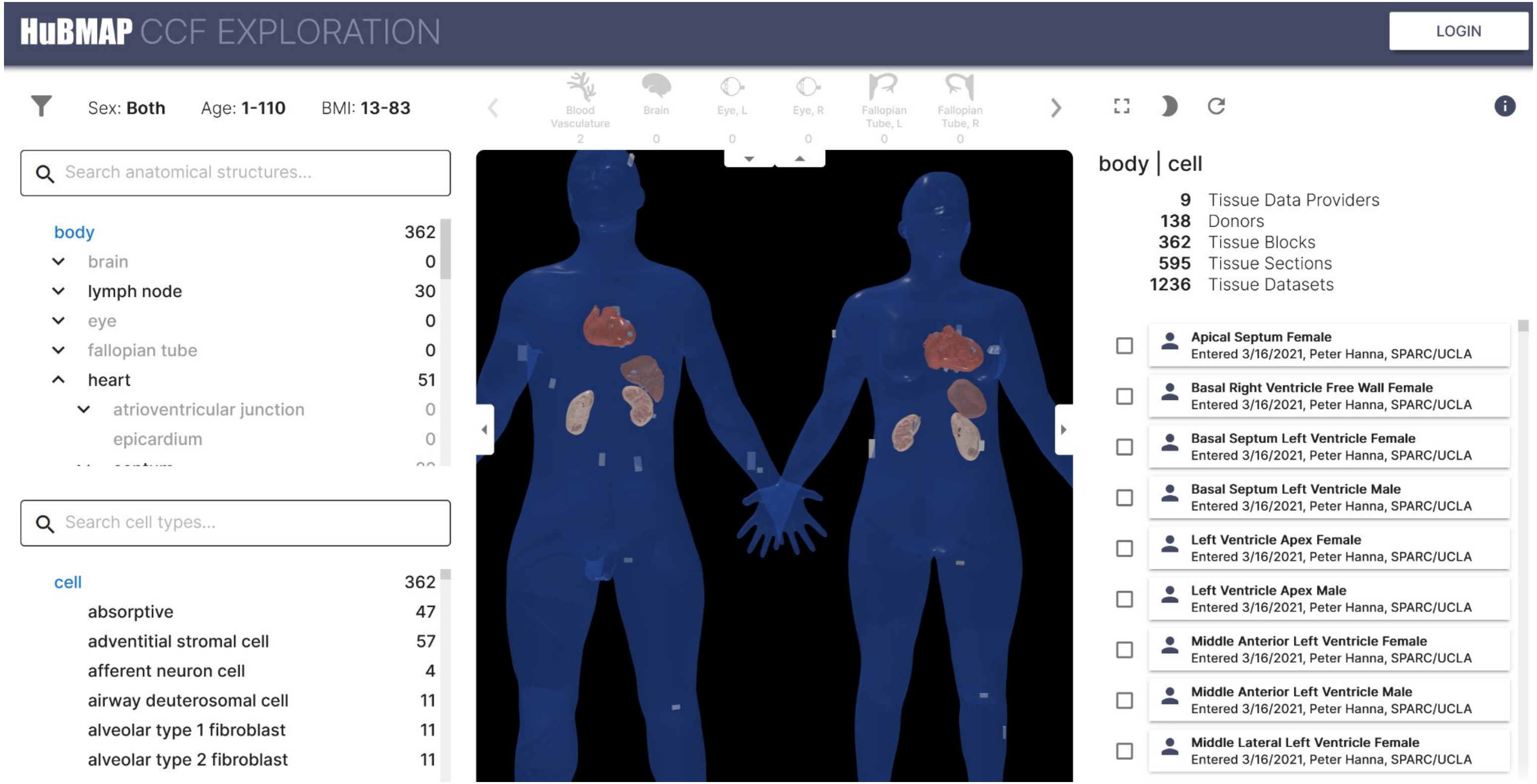
+



https://hubmapconsortium.github.io/ccf/
https://www.nature.com/articles/s41556-021-00788-6

https://cns-iu.github.io/spoke-vis/home
https://onlinelibrary.wiley.com/doi/10.1002/aaai.12037

Question: How to bidrectionally link ontologies to 1/2/3D references



Sex: **Both** Age: **1-110** BMI: **13-83**

Search anatomical structures...

body	362
brain	0
lymph node	30
eye	0
fallopian tube	0
heart	51
atrioventricular junction	0
epicardium	0

Search cell types...

cell	362
absorptive	47
adventitial stromal cell	57
afferent neuron cell	4
airway deuterosomal cell	11
alveolar type 1 fibroblast	11
alveolar type 2 fibroblast	11

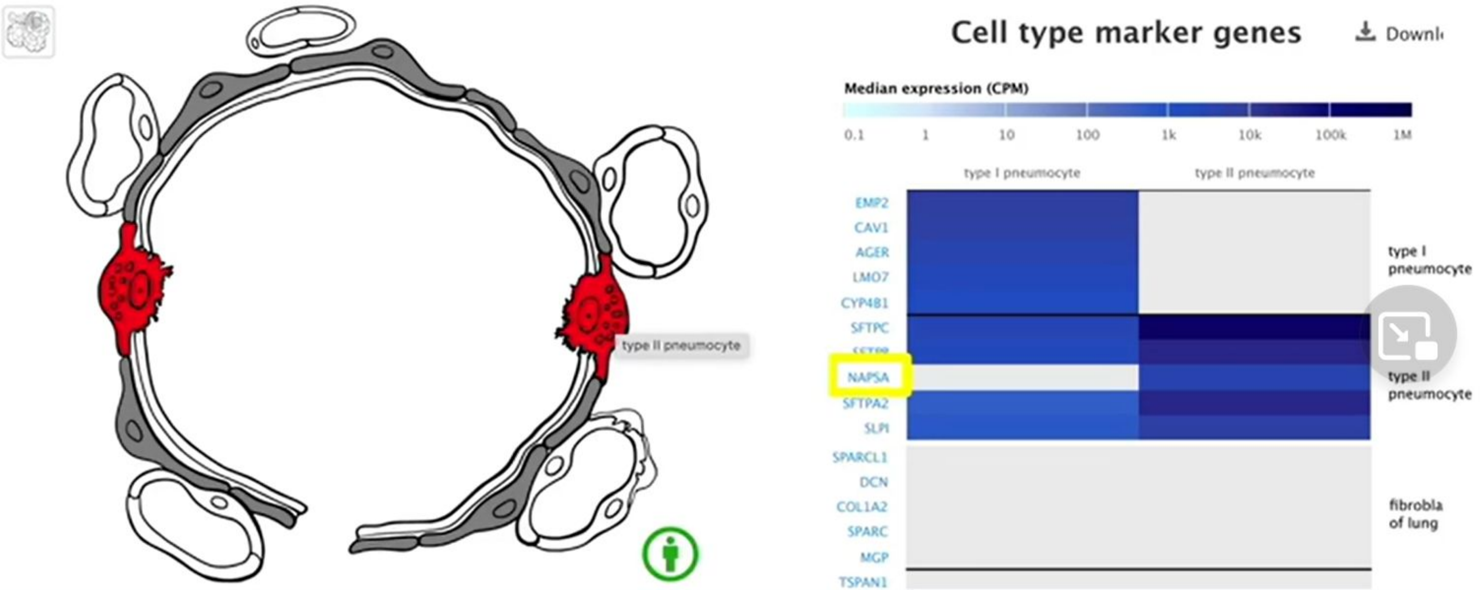
body | cell

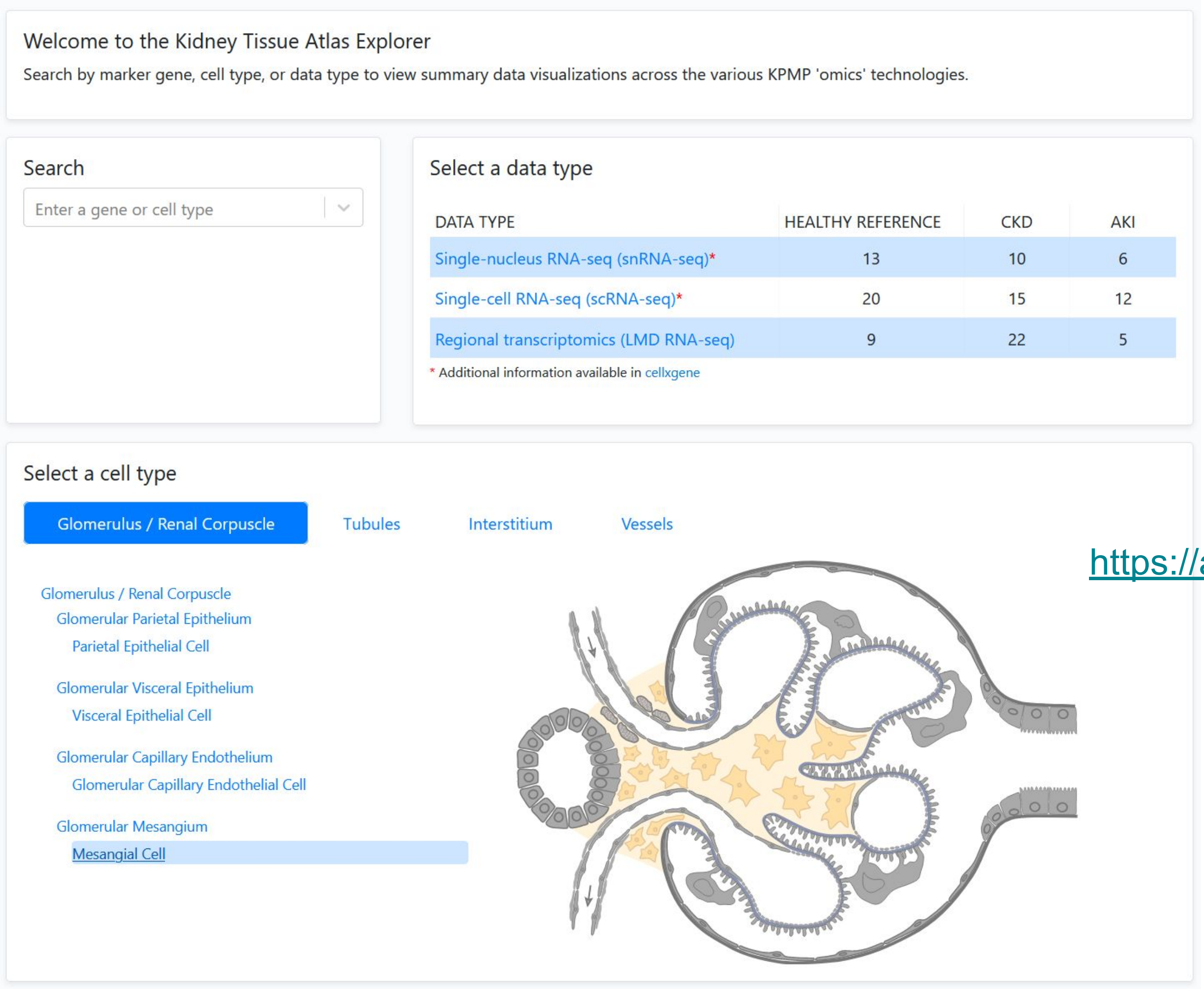
9	Tissue Data Providers
138	Donors
362	Tissue Blocks
595	Tissue Sections
1236	Tissue Datasets

- ☐ **Apical Septum Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Basal Right Ventricle Free Wall Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Basal Septum Left Ventricle Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Basal Septum Left Ventricle Male** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Left Ventricle Apex Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Left Ventricle Apex Male** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Middle Anterior Left Ventricle Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Middle Anterior Left Ventricle Male** Entered 3/16/2021, Peter Hanna, SPARC/UCLA
- ☐ **Middle Lateral Left Ventricle Female** Entered 3/16/2021, Peter Hanna, SPARC/UCLA

https://portal.hubmapconsortium.org/ccf-eui

NAPSA marker gene in type II pneumocytes





Welcome to the Kidney Tissue Atlas Explorer

Search by marker gene, cell type, or data type to view summary data visualizations across the various KPMP 'omics' technologies.

Search

Select a data type

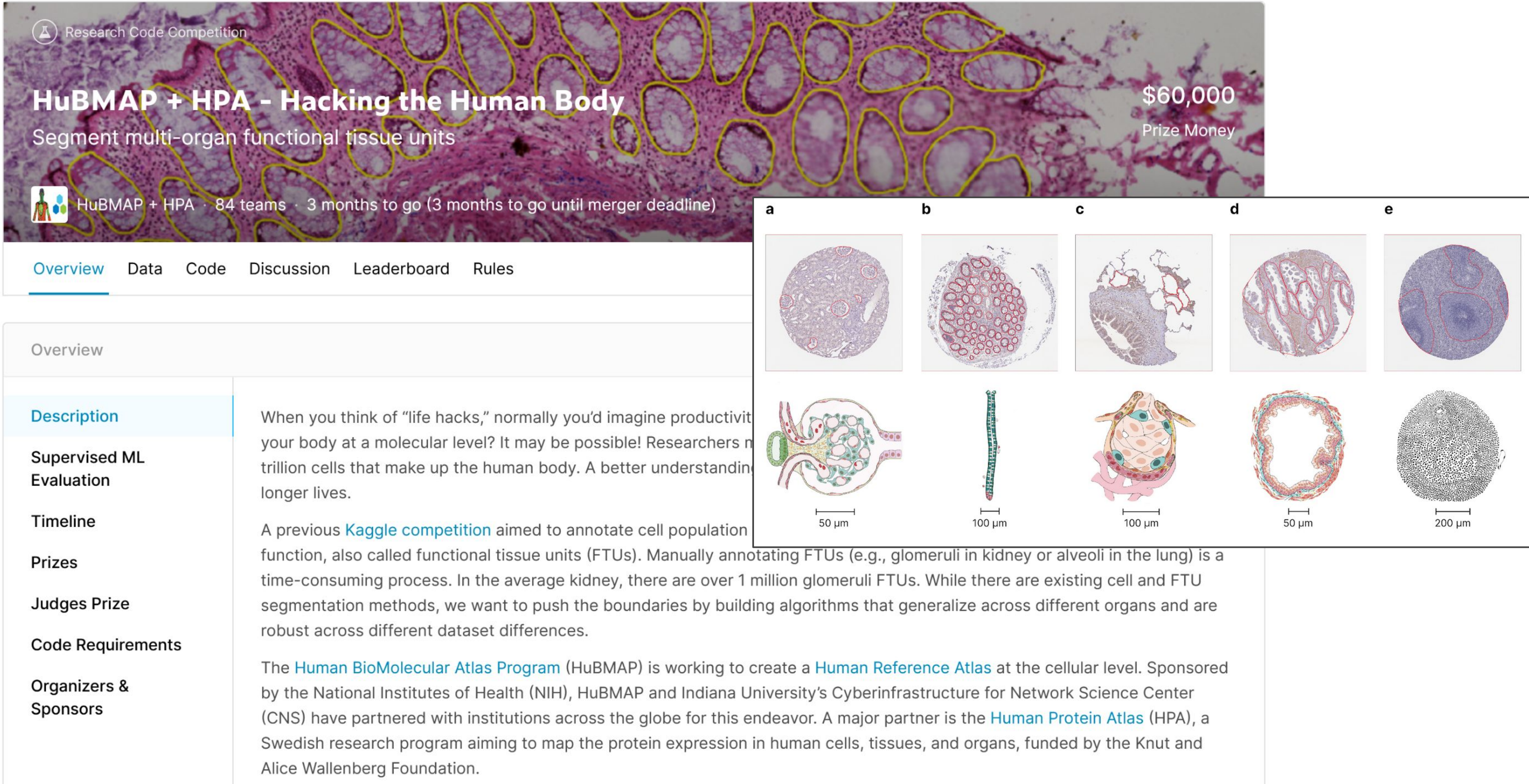
DATA TYPE	HEALTHY REFERENCE	CKD	AKI
Single-nucleus RNA-seq (snRNA-seq)*	13	10	6
Single-cell RNA-seq (scRNA-seq)*	20	15	12
Regional transcriptomics (LMD RNA-seq)	9	22	5

* Additional information available in cellxgene

Select a cell type

- Glomerulus / Renal Corpuscle
  - Glomerular Parietal Epithelium
    - Parietal Epithelial Cell
  - Glomerular Visceral Epithelium
    - Visceral Epithelial Cell
  - Glomerular Capillary Endothelium
    - Glomerular Capillary Endothelial Cell
  - Glomerular Mesangium
    - Mesangial Cell

https://atlas.kpmp.org/explorer/



Research Code Competition

HuBMAP + HPA - Hacking the Human Body

Segment multi-organ functional tissue units

$60,000
Prize Money

HuBMAP + HPA · 84 teams · 3 months to go (3 months to go until merger deadline)

Overview Data Code Discussion Leaderboard Rules

Overview

- Description
- Supervised ML Evaluation
- Timeline
- Prizes
- Judges Prize
- Code Requirements
- Organizers & Sponsors

When you think of “life hacks,” normally you’d imagine productivit
your body at a molecular level? It may be possible! Researchers m
trillion cells that make up the human body. A better understandin
longer lives.

A previous Kaggle competition aimed to annotate cell population
function, also called functional tissue units (FTUs). Manually annotating FTUs (e.g., glomeruli in kidney or alveoli in the lung) is a time-consuming process. In the average kidney, there are over 1 million glomeruli FTUs. While there are existing cell and FTU segmentation methods, we want to push the boundaries by building algorithms that generalize across different organs and are robust across different dataset differences.

The Human BioMolecular Atlas Program (HuBMAP) is working to create a Human Reference Atlas at the cellular level. Sponsored by the National Institutes of Health (NIH), HuBMAP and Indiana University’s Cyberinfrastructure for Network Science Center (CNS) have partnered with institutions across the globe for this endeavor. A major partner is the Human Protein Atlas (HPA), a Swedish research program aiming to map the protein expression in human cells, tissues, and organs, funded by the Knut and Alice Wallenberg Foundation.

https://www.kaggle.com/competitions/hubmap-organ-segmentation

Flash Talks

Flash Talks by Experts

Other experts in the room
Tony Burdett, EBI, UK
Paulo Czarnewski, Scilifelab, SE

Melissa Clarkson, UKY, USA

Wei Kheng Teh, EBI, UK

Jason Hilton, Stanford U, USA

Evan Biederstedt, Harvard Medical School, USA

Angela Pisco, CZI BioHub, USA

Fabian Theis, Helmholtz-Muenchen, Germany

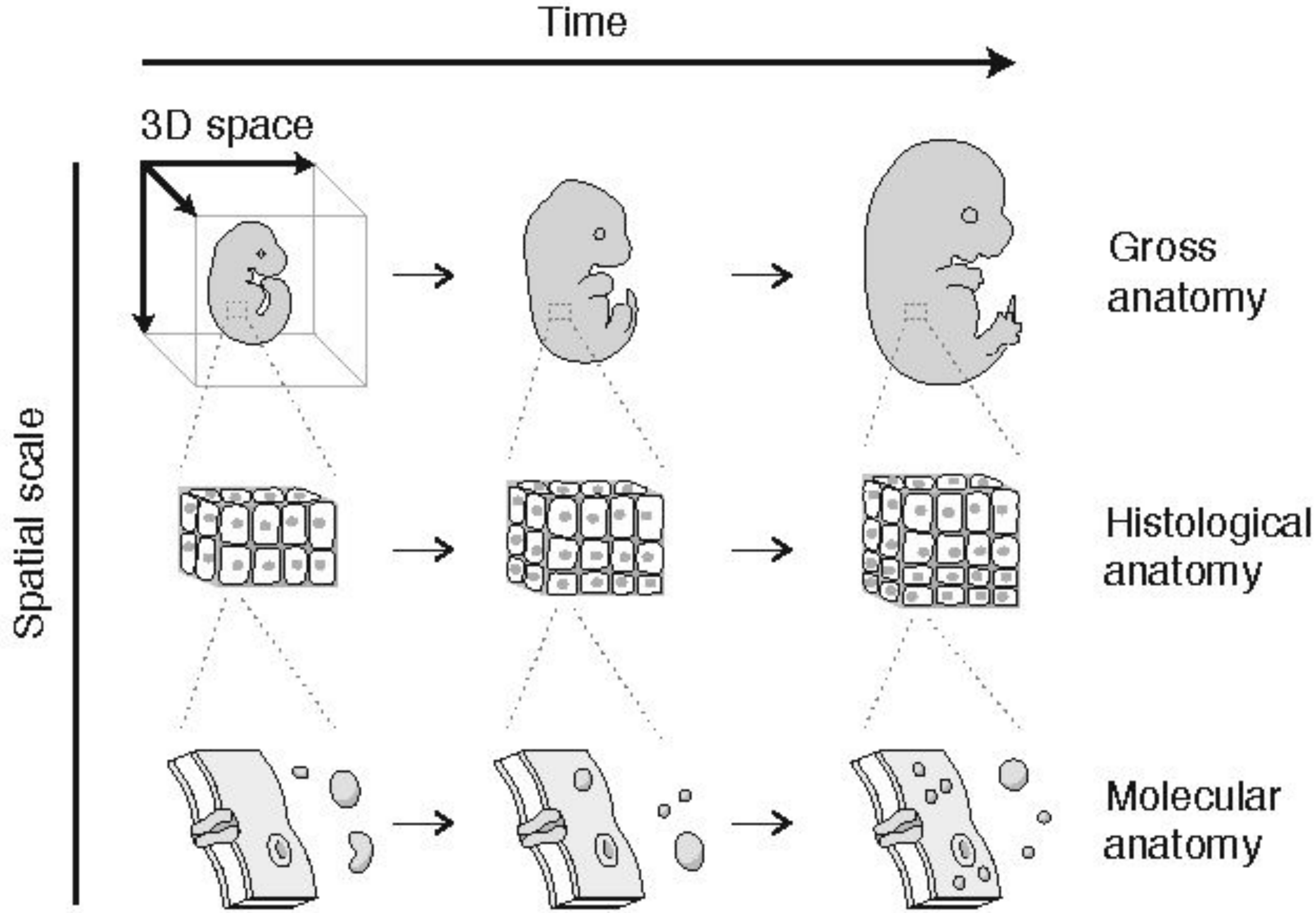
Bruce W. Herr II, Indiana University, USA

Melissa Clarkson

Melissa Clarkson, UKY, USA

Representing human anatomy at the scale of organs to tissues —

The Foundational Model of Human Anatomy project


Time
3D space
Spatial scale
Gross anatomy
Histological anatomy
Molecular anatomy

The Foundational Model of Anatomy (FMA) is a reference ontology for adult canonical anatomy

- A project of the Structural Informatics Group at the University of Washington
- Modeled in OWL
- Over 100,000 anatomical structures represented as classes
- Over 100 types of relations among classes

The FMA has a number of issues that affect its ability to continue to serve as a knowledge base

- Variations in modeling schemes for similar anatomy in different parts of the body
- Incomplete content
- Not easy to understand meaning of some classes

I am beginning a project to develop a derivative of the FMA that will eventually replace the FMA

Foundational Model of Human Anatomy (FMHA)

Development strategy will improve:

- consistency of modeling
- completeness
- clarity

Consistency of modeling will be improved by using patterns

Example of inconsistencies in the FMA…

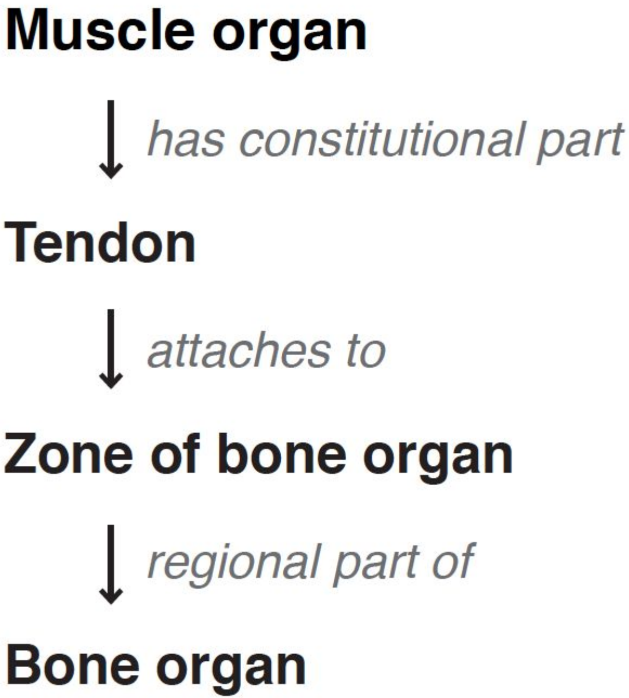
Question:
How is a muscle related to the bone it attaches to?

Consistency of modeling will be improved by using patterns

Example of inconsistencies in the FMA

Subject superclass	**Relation**	**Object superclass**	**Number of axioms**
Tendon	attaches to	Zone of bone organ	207
Zone of muscle organ	has insertion	Zone of bone organ	60
Muscle organ	has insertion	Zone of bone organ	36
Muscle organ	attaches to	Zone of bone organ	11

Consistency of modeling will be improved by using patterns



Muscle organ

has constitutional part

Tendon

attaches to

Zone of bone organ

regional part of

Bone organ

The vision:

The Foundational Model of Human Anatomy (FMHA) ontology will be a computable representation of human anatomy, linked to graphics with computer-readable semantics — creating an “illustrated ontology”.

Wei Kheng Teh

Wei Kheng Teh, EBI, UK

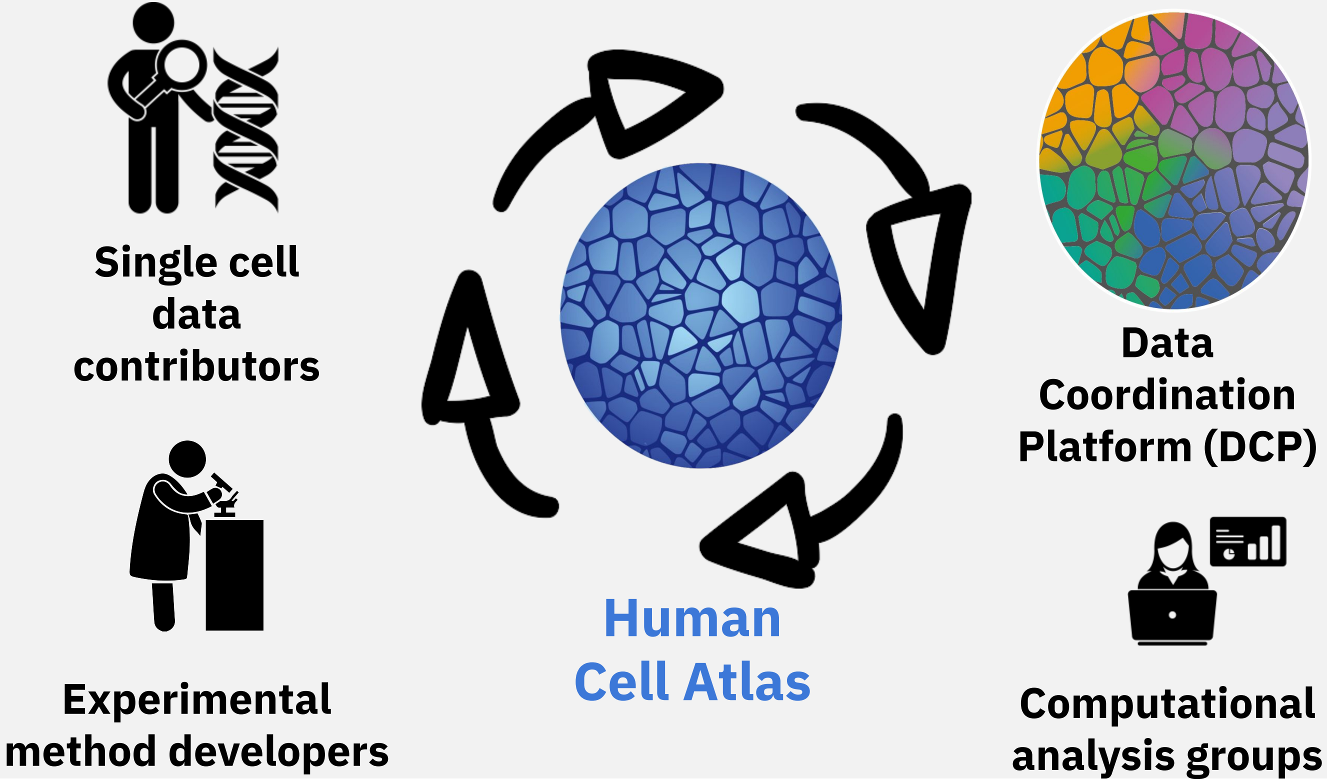
HUMAN CELL ATLAS
DATA PORTAL

https://data.humancellatlas.org/

Cell Ontologies, Annotation & Metadata

How can we make ontology annotation following standard schemas easy, efficient, and accurate while leaving room for revising and adding to existing ontologies?

The Human Cell Atlas Data Coordination Platform (DCP)



Single cell data contributors

Data Coordination Platform (DCP)

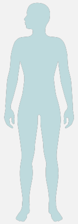
Human Cell Atlas

Experimental method developers

Computational analysis groups

HCA - example of desired metadata

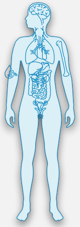
Donor



NCBI TAXON ID
GENUS SPECIES
BIOLOGICAL SEX
ALIVE AT COLLECTION

KNOWN DISEASE(S)
ETHNICITY

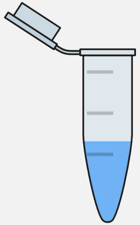
Specimen from organism



NCBI TAXON ID
GENUS SPECIES
ORGAN
INSDC EXPERIMENT ACCESSION

ORGAN PART
KNOWN DISEASE(S)
POST-MORTEM INTERVAL
PRESERVATION METHOD

Cell Suspension

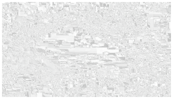


PERCENT CELL VIABILITY
CELL VIABILITY METHOD
CELL VIABILITY RESULT
PERCENT NECROTIC CELLS

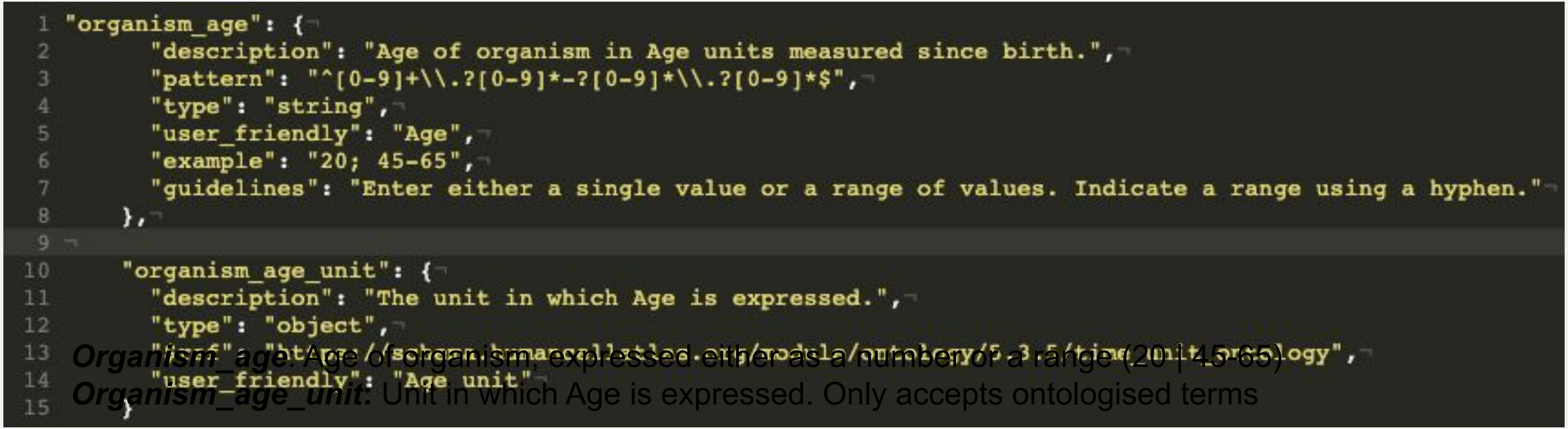
Sequencing

CELL BARCODE (3)
INPUT NUCLEIC ACID MOLECULE
NUCLEIC ACID SOURCE
LIBRARY CONSTRUCTION METHOD
END BIAS
STRAND
UMI BARCODE (3)
INSTRUMENT MANUFACTURER AND MODEL
PAIRED END
SEQUENCING METHOD

The HCA Metadata Schema



Metadata schemas contain all information required to **understand and interpret** the standard



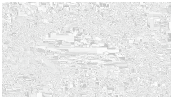
```
"organism_age": {
        "description": "Age of organism in Age units measured since birth.",
        "pattern": "^[0-9]+\\.?[0-9]*-?[0-9]*\\.?[0-9]*$",
        "type": "string",
        "user_friendly": "Age",
        "example": "20; 45-65",
        "guidelines": "Enter either a single value or a range of values. Indicate a range using a hyphen."
    },

    "organism_age_unit": {
        "description": "The unit in which Age is expressed.",
        "type": "object",
        "$ref": "https://schema.humancellatlas.org/module/ontology/5.3.5/time_unit_ontology",
        "user_friendly": "Age unit"
    }
```

Organism_age: Age of organism, expressed either as a number or a range (20 | 45-65)

Organism_age_unit: Unit in which Age is expressed. Only accepts ontologised terms

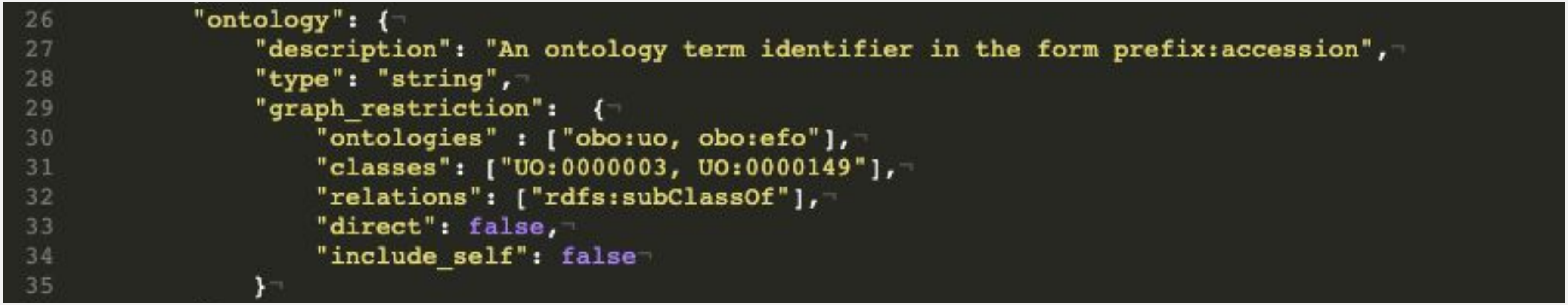
The HCA Metadata Schema



The schema can be expanded to validate for ontologies

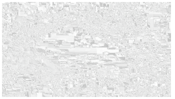
Any term accepted under classes
UO:0000003 (time unit)
UO:0000149 (derived time unit)

Any other term not descending from these 2 is **rejected** in this field



```
"ontology": {
    "description": "An ontology term identifier in the form prefix:accession",
    "type": "string",
    "graph_restriction":  {
        "ontologies" : ["obo:uo, obo:efo"],
        "classes": ["UO:0000003, UO:0000149"],
        "relations": ["rdfs:subClassOf"],
        "direct": false,
        "include_self": false
    }
```

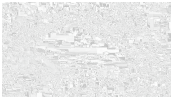
The HCA Metadata Schema

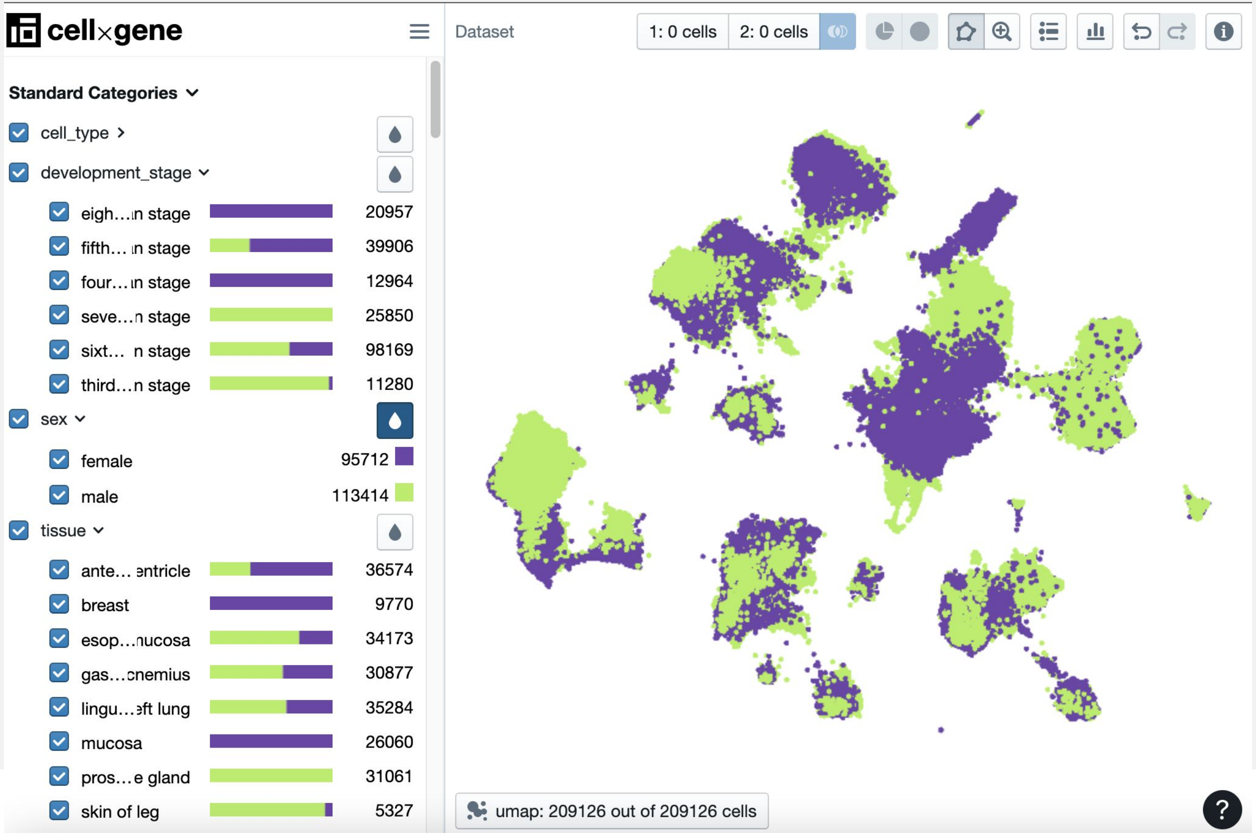


Ethnicity
Developmental Stage
Age Unit
Tissue
Organ
Known Diseases
Library Preparation Methods
Sequencing Machine
Cell Cycle
…

OLS
ONTOLOGY SEARCH

Cell Ontologies and Annotation





Standard Categories

☑ cell_type

☑ development_stage

☑ eigh…n stage	20957
☑ fifth… n stage	39906
☑ four…n stage	12964
☑ seve…n stage	25850
☑ sixt… n stage	98169
☑ third…n stage	11280

☑ sex

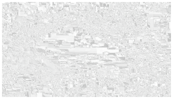
☑ female	95712 🟪
☑ male	113414 🟩

☑ tissue

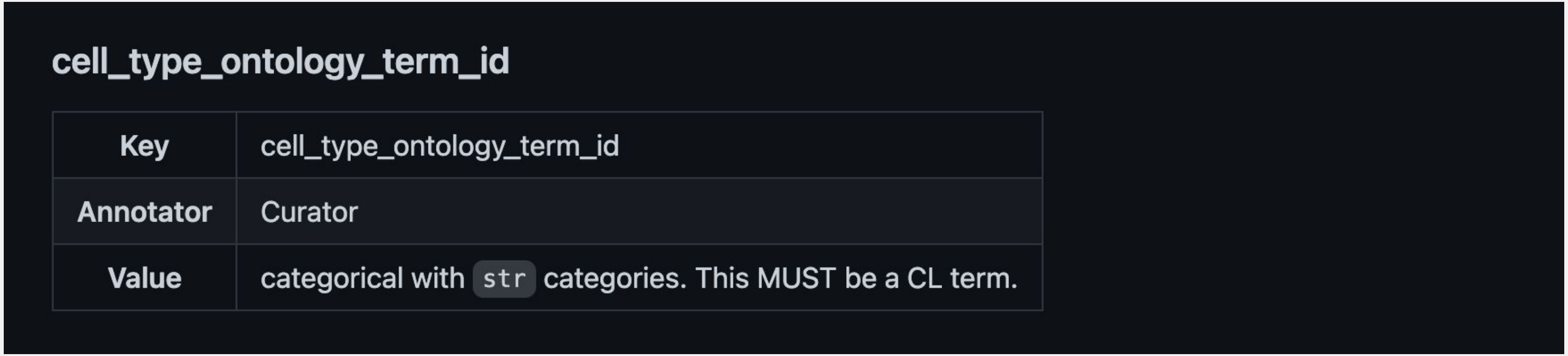
☑ ante… entricle	36574
☑ breast	9770
☑ esop…nucosa	34173
☑ gas…cnemius	30877
☑ lingu…eft lung	35284
☑ mucosa	26060
☑ pros…e gland	31061
☑ skin of leg	5327

umap: 209126 out of 209126 cells

Cell Ontologies and Annotation



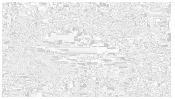
CellxGene - cell_type_ontology_term_id



cell_type_ontology_term_id

Key	cell_type_ontology_term_id
Annotator	Curator
Value	categorical with `str` categories. This MUST be a CL term.

Challenges and Future Development



- Novel Cell Types and adding new terms to CL
- Releasing annotations with data
- Talk to us here or at wrangler-team@data.humancellatlas.org

Thanks to our partners

EMBL-EBI

BROAD INSTITUTE

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Programs and Funders

- Related and complementary initiatives
- Diverse funded projects across the globe
- Support for central efforts: DCP, meetings, ethics, equity

CZ

wellcome

Horizon 2020 European Union Funding for Research & Innovation

NIH

BILL & MELINDA GATES foundation

Inserm

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

The Manton Foundation

MRC Medical Research Council

Knut and Alice Wallenberg Foundation

THE KAVLI FOUNDATION

FET FLAGSHIPS

Jason Hilton

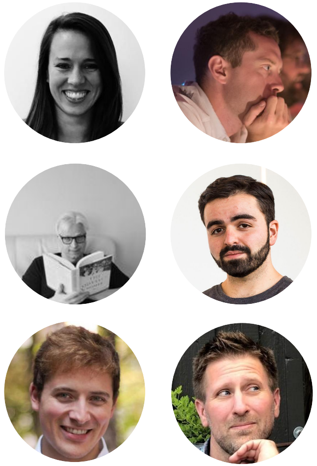
Jason Hilton, Stanford U, USA

cell×gene Data Portal

cellxgene.cziscience.com

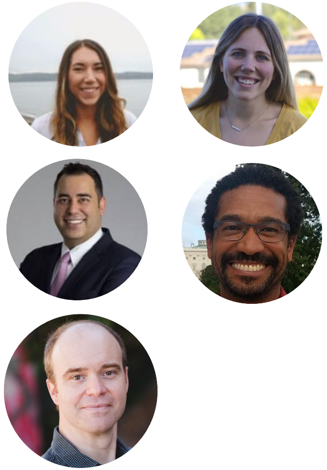
CZI Single-Cell Team

Product & Design

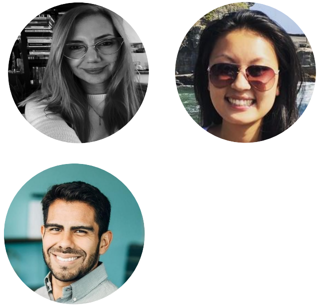


Engineering

Science Program



Data & Statistics



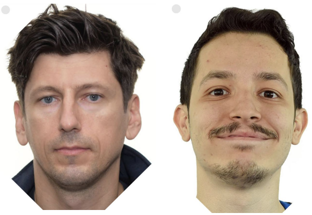
Clever Canary

Curation Teams

Lattice, Stanford Univ.

HCA DCP, UCSC & EBI

Sanger

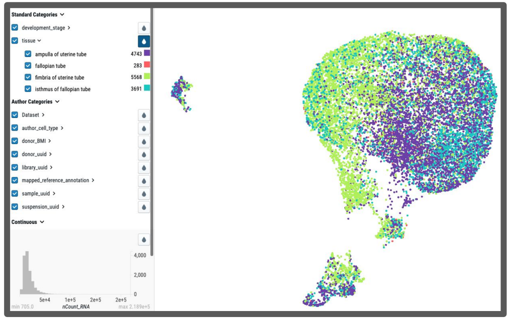


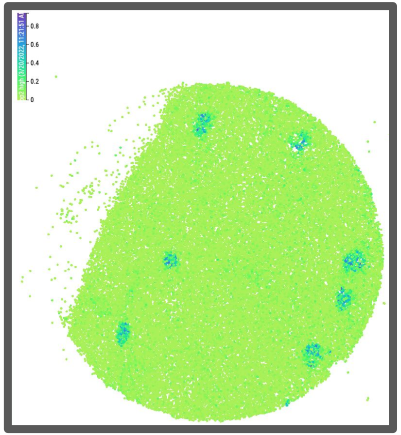
cellxgene Data Portal

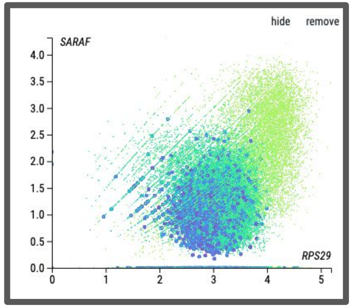
cellxgene.cziscience.com

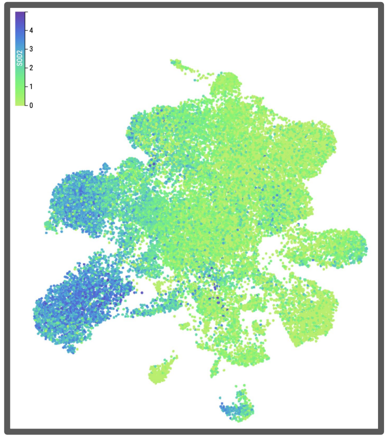
A. Publish single collection of data

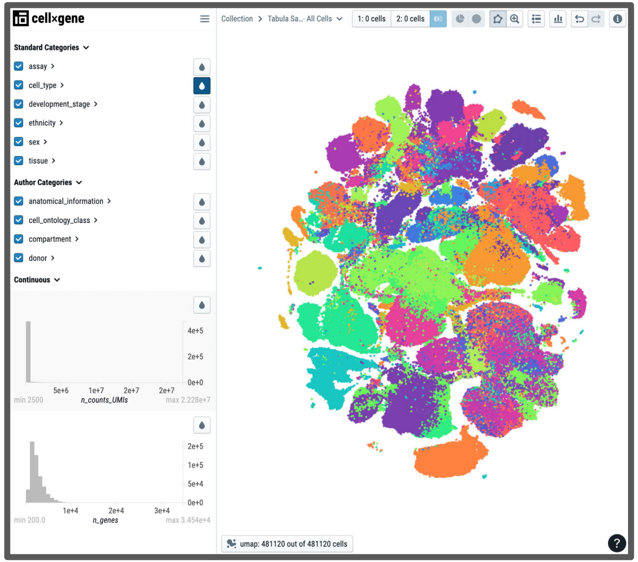
Visualize and download data











cellxgene Data Portal

cellxgene.cziscience.com

A. Publish single collection of data

Visualize and download data

B. Enable cross-collection integration & other reuse cases

Requires some standardization

cellxgene Data Portal

Standards: schema 2.0.0

- All data available as AnnData (.h5ad) & Seurat (.rds)

cellxgene Data Portal

Standards: schema 2.0.0

- AnnData & Seurat
- raw counts are required
    - unscaled, pre-normalization counts per cell
    - enable integration without realignment

cellxgene Data Portal

Standards: schema 2.0.0

- AnnData & Seurat
- raw counts
- feature metadata
  - Ensembl IDs are required
  - stable identifiers, as opposed to symbols

cellxgene Data Portal

Standards: schema 2.0.0

- AnnData & Seurat
- raw counts
- feature metadata
  - Ensembl IDs are required
  - stable identifiers, as opposed to symbols

Challenge

accurately mapping from gene symbols to Ensembl IDs

cellxgene Data Portal

Standards: schema 2.0.0

- AnnData & Seurat
- raw counts
- feature metadata
- observation metadata
  - annotated to the most specific available ontology term

Required field	Ontology
organism	**NCBITaxon**
donor_id	
development_stage	**HsapDv/MmusDv**
sex	**PATO**
ethnicity	
disease	**MONDO**
tissue	**UBERON**
cell_type	**CL**
assay	**EFO**
observation_type	**[cell,nucleus]**

cellxgene Data Portal

Standards: schema 2.0.0

Challenge

How to standardize ethnicity

- AnnData & Seurat
- raw counts
- feature metadata
- observation metadata
  - annotated to the most specific available ontology term

Required field	Ontology
organism	**NCBITaxon**
donor_id	
development_stage	**HsapDv/MmusDv**
sex	**PATO**
ethnicity	
disease	**MONDO**
tissue	**UBERON**
cell_type	**CL**
assay	**EFO**
observation_type	**[cell,nucleus]**

cellxgene Data Portal

Standards: schema 2.0.0

- AnnData & Seurat
- raw counts
- feature metadata
- observation metadata
  - annotated to the most specific available ontology term

Challenge

How to standardize ethnicity

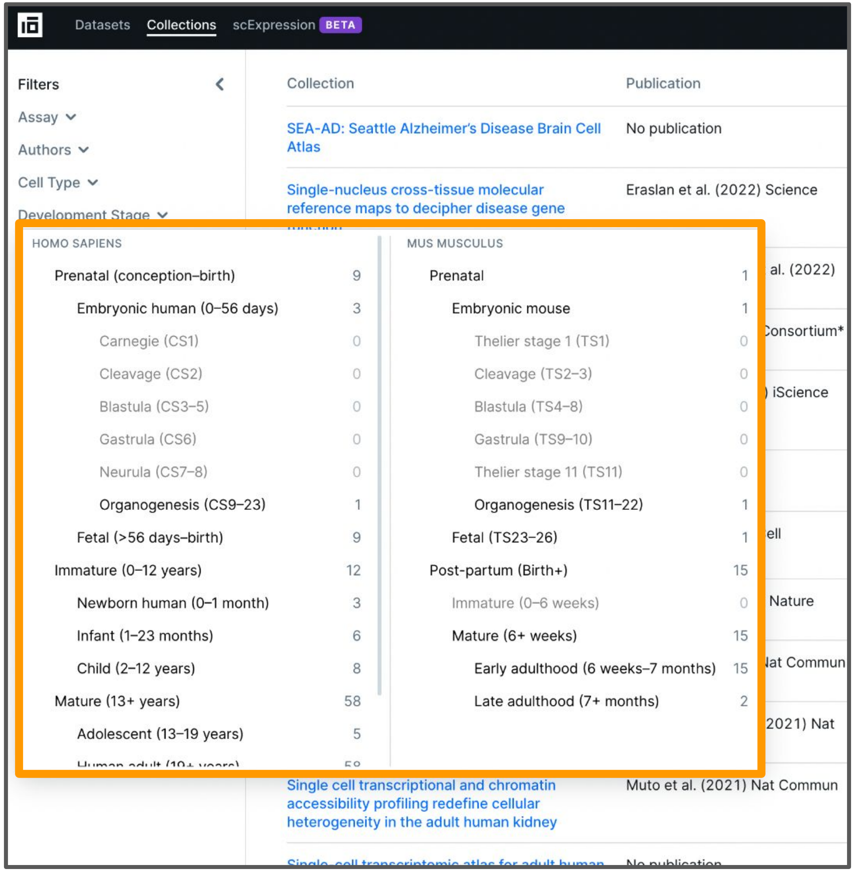
Required field	Ontology
organism	**NCBITaxon**
donor_id	
development_stage	**HsapDv/MmusDv**
sex	**PATO**
ethnicity	
disease	**MONDO**
tissue	**UBERON**
cell_type	**CL**
assay	**EFO**
observation_type	**[cell,nucleus]**

Challenge

How to standardize unannotated cells

cellxgene Data Portal

Ontology usage



Filters

Assay

Authors

Cell Type

Development Stage

Collection	Publication
SEA-AD: Seattle Alzheimer's Disease Brain Cell Atlas	No publication
Single-nucleus cross-tissue molecular reference maps to decipher disease gene function	Eraslan et al. (2022) Science
	al. (2022)
	Consortium*
	) iScience
	ell
	Nature
	Nat Commun
	2021) Nat
Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney	Muto et al. (2021) Nat Commun
Single-cell transcriptomic atlas for adult human	No publication

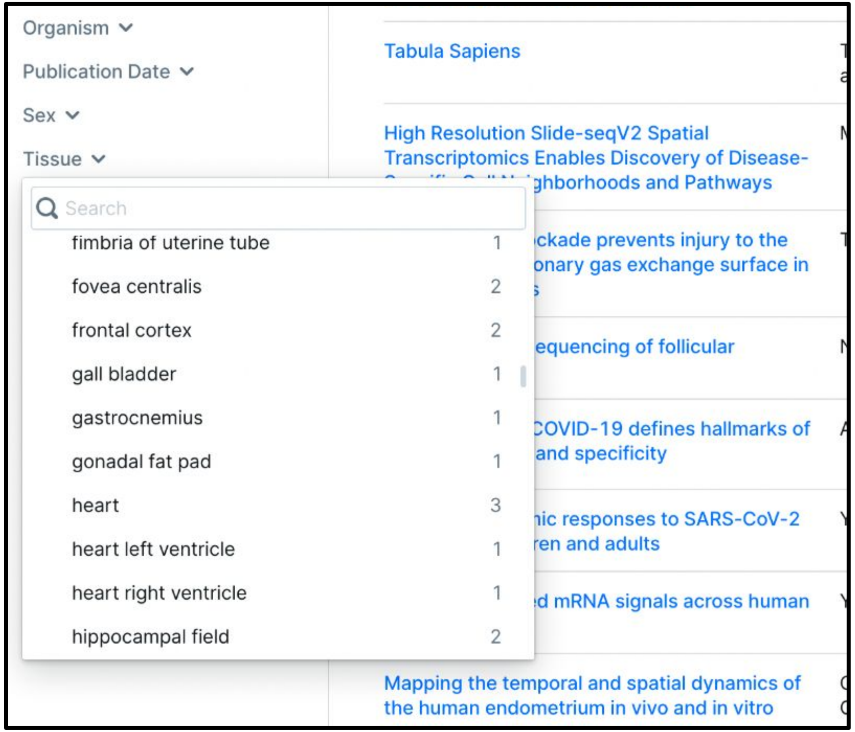
HOMO SAPIENS	
Prenatal (conception–birth)	9
Embryonic human (0–56 days)	3
Carnegie (CS1)	0
Cleavage (CS2)	0
Blastula (CS3–5)	0
Gastrula (CS6)	0
Neurula (CS7–8)	0
Organogenesis (CS9–23)	1
Fetal (>56 days–birth)	9
Immature (0–12 years)	12
Newborn human (0–1 month)	3
Infant (1–23 months)	6
Child (2–12 years)	8
Mature (13+ years)	58
Adolescent (13–19 years)	5
Human adult (19+ years)	58

MUS MUSCULUS	
Prenatal	1
Embryonic mouse	1
Thelier stage 1 (TS1)	0
Cleavage (TS2–3)	0
Blastula (TS4–8)	0
Gastrula (TS9–10)	0
Thelier stage 11 (TS11)	0
Organogenesis (TS11–22)	1
Fetal (TS23–26)	1
Post-partum (Birth+)	15
Immature (0–6 weeks)	0
Mature (6+ weeks)	15
Early adulthood (6 weeks–7 months)	15
Late adulthood (7+ months)	2

Group data by higher-level terms

cellxgene Data Portal

Ontology usage



Organism

Publication Date

Sex

Tissue

Search

fimbria of uterine tube	1
fovea centralis	2
frontal cortex	2
gall bladder	1
gastrocnemius	1
gonadal fat pad	1
heart	3
heart left ventricle	1
heart right ventricle	1
hippocampal field	2

Tabula Sapiens

High Resolution Slide-seqV2 Spatial Transcriptomics Enables Discovery of Disease-
ghborhoods and Pathways

ckade prevents injury to the
onary gas exchange surface in
s

equencing of follicular

COVID-19 defines hallmarks of
and specificity

nic responses to SARS-CoV-2
ren and adults

d mRNA signals across human

Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro

Group data by higher-level terms

Grouping tissue

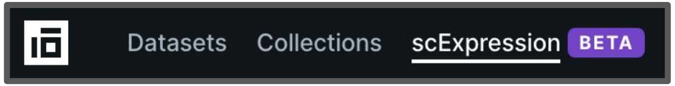
“system” and/or “organ”

Challenge

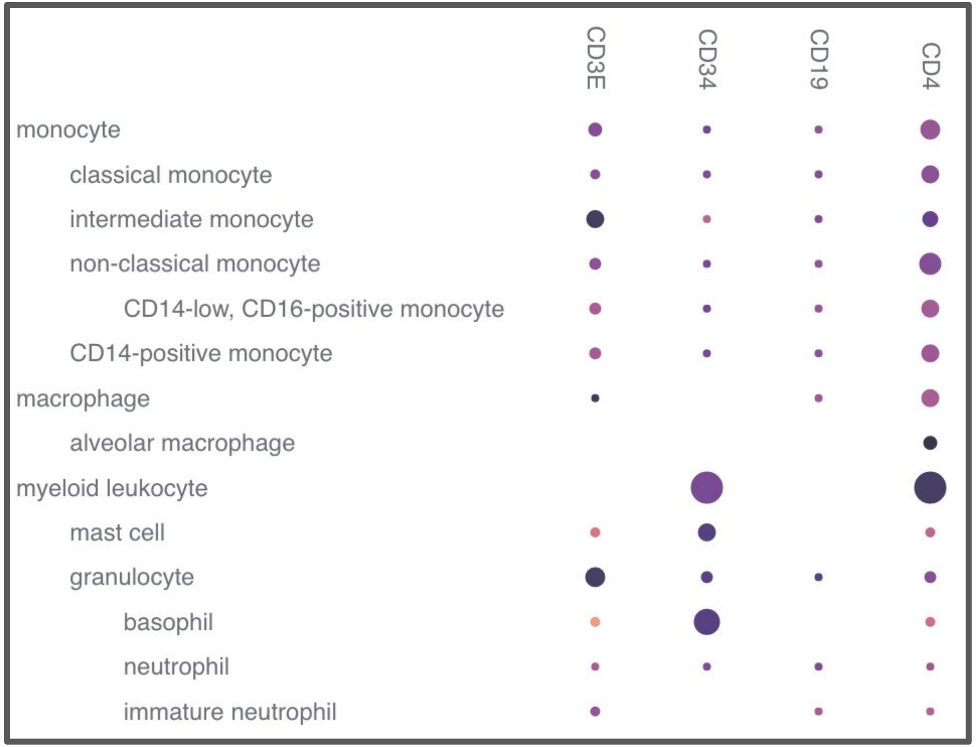
How to group by higher-level terms

cellxgene Data Portal

Ontology usage



Ontology-aware organization of per-cell expression



cellxgene Data Portal

Challenges

- Mapping Ensembl IDs from gene symbols
- Standardizing ethnicity
- Standardizing unannotated cells
- Understanding how users want to find & filter data

THANK YOU!

Evan Biederstedt

Evan Biederstedt, Harvard Medical School, USA

https://speakerdeck.com/evanbiederstedt/hca-general-meeting-2022-cell-annotation-platform

https://rc1.celltype.info - Demo time! (**Update:** on YouTube soon)

9:00 AM

9:00 AM - 10:30 AM GMT +2 / 3:00 AM - 4:30 AM Your local time (1 Hour, 30 Min)

Plenary Session 4 - Opportunities and Challenges Using and Accessing Data

Palais Niederosterreich, Main Hall

Session Chair, Leyla Tarhan

1. Nils Gehlenborg, "Atlas User Interfaces to Accelerate Data Access"

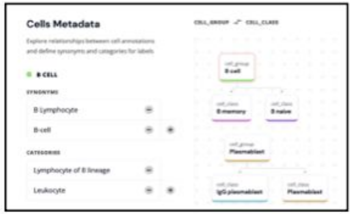
Irene Papatheodorou
Team Leader
EMBL-EBI
Speaker

Malte Luecken
Postdoctoral Researcher
Helmholtz Munich
Speaker

Evan Biederstedt
Research Scientist
Harvard Medical School
Speaker

Annotation Suggestions in Real-time

Text-based



Cells Metadata

Explore relationships between cell annotations and define synonyms and categories for labels

B CELL

SYNONYMS

B Lymphocyte

B-cell

CATEGORIES

Lymphocyte of B lineage

Leukocyte

CELL_GROUP CELL_CLASS

cell_group B cell

cell_class B memory

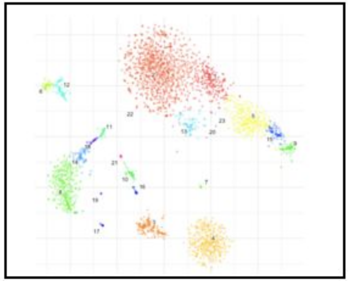
cell_class B naive

cell_group Plasmablast

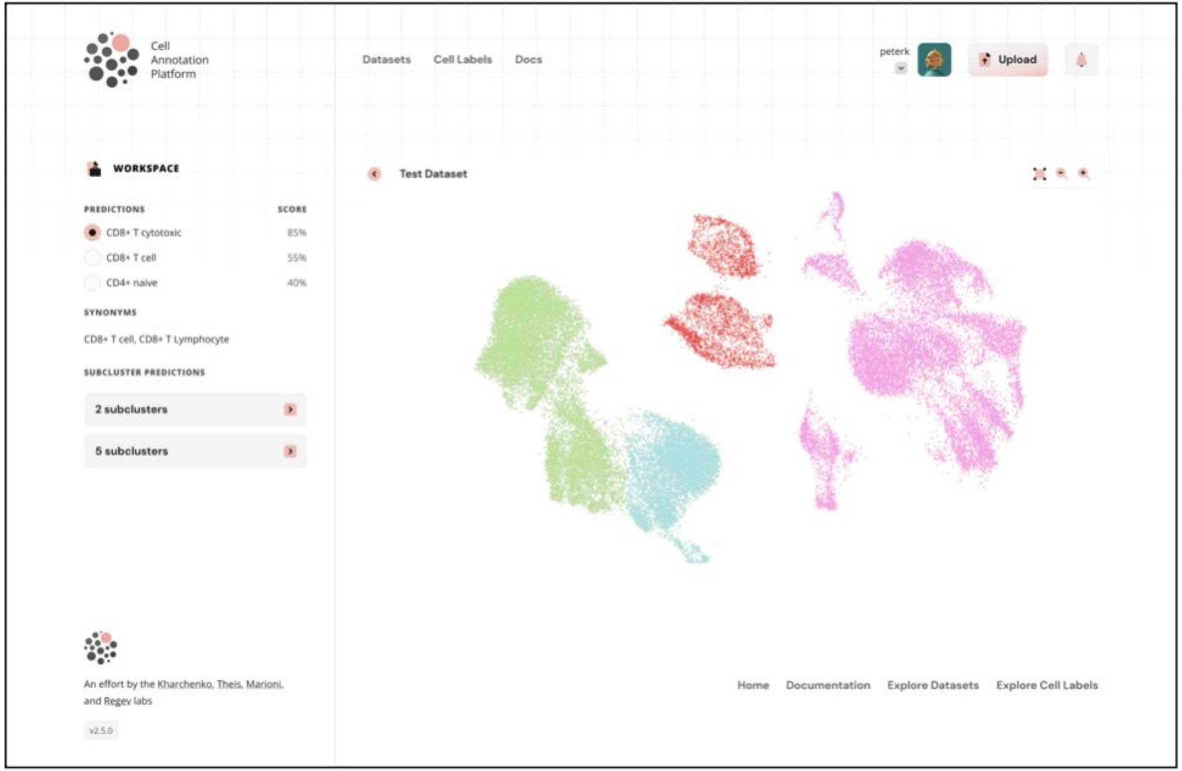
cell_class IgG plasmablast

cell_class Plasmablast

Molecular-based



→



Cell Annotation Platform

Datasets Cell Labels Docs

peterk Upload

WORKSPACE

Test Dataset

PREDICTIONS	SCORE
● CDB+ T cytotoxic	85%
○ CD8+ T cell	55%
○ CD4+ naive	40%

SYNONYMS

CD8+ T cell, CD8+ T Lymphocyte

SUBCLUSTER PREDICTIONS

2 subclusters

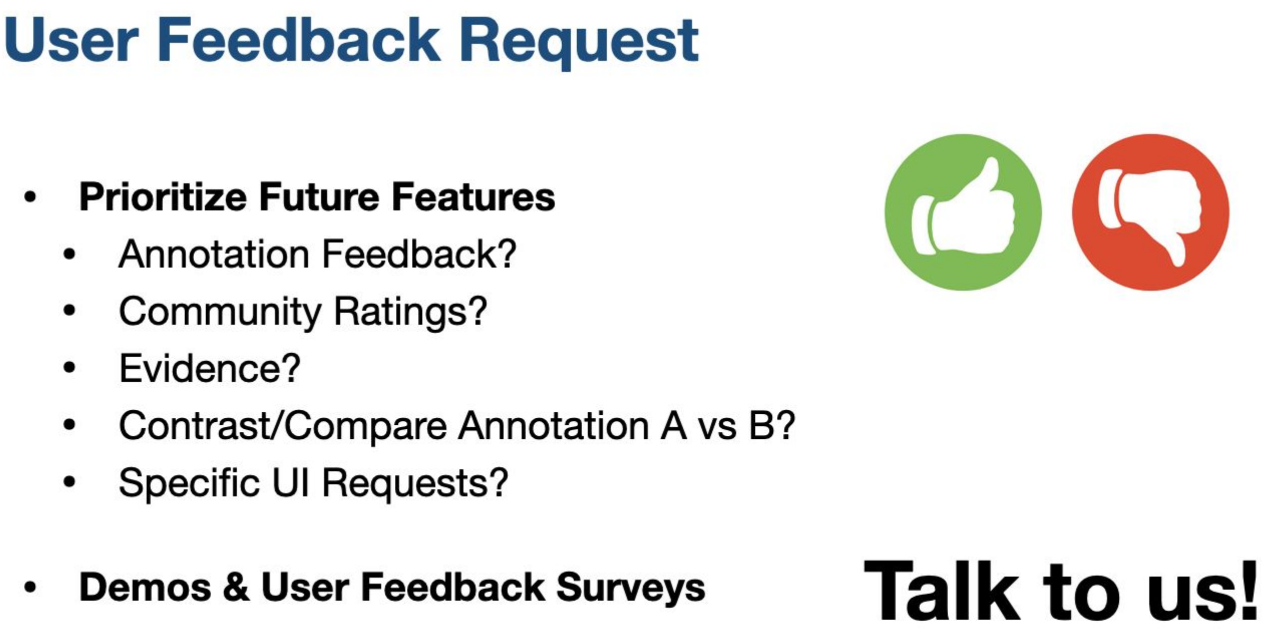
5 subclusters

An effort by the Kharchenko, Theis, Marioni, and Regev labs

v2.5.0

Home Documentation Explore Datasets Explore Cell Labels

Evan Biederstedt, Harvard Medical School, USA



User Feedback Request

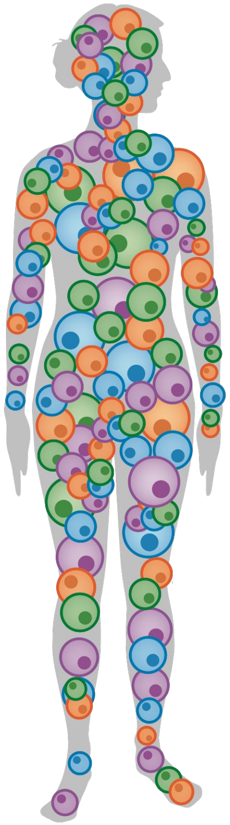
- **Prioritize Future Features**
  - Annotation Feedback?
  - Community Ratings?
  - Evidence?
  - Contrast/Compare Annotation A vs B?
  - Specific UI Requests?
- **Demos & User Feedback Surveys**

Talk to us!

Angela Pisco

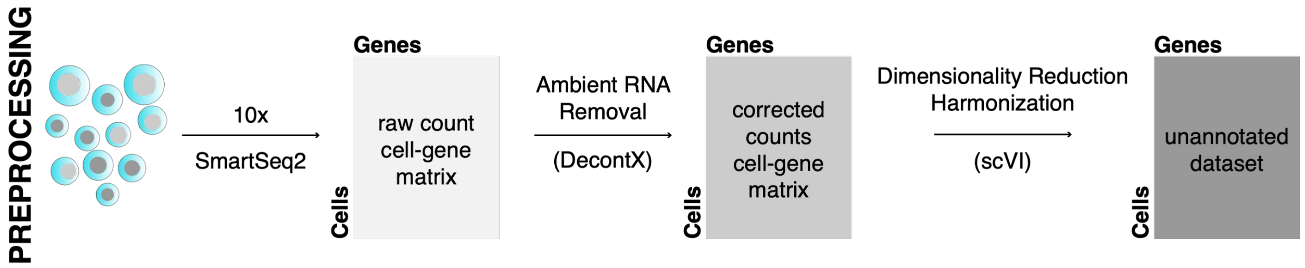
Angela Pisco, CZ Biohub, USA

Tools for Automating Cell Type Calling



angela.pisco@czbiohub.org @drAOPisco @aopisco

CELL TYPE ANNOTATION WHEN BUILDING AN ATLAS



PREPROCESSING

10x → SmartSeq2

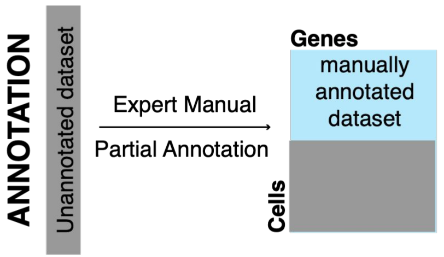
Genes / **Cells**: raw count cell-gene matrix

Ambient RNA Removal → (DecontX)

Genes / **Cells**: corrected counts cell-gene matrix

Dimensionality Reduction Harmonization → (scVI)

Genes / **Cells**: unannotated dataset



ANNOTATION

Unannotated dataset

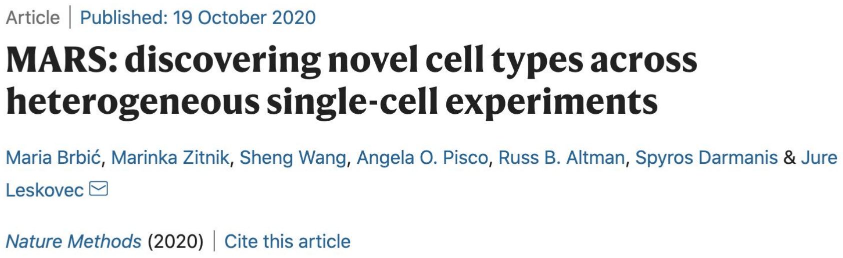
Expert Manual → Partial Annotation

Genes / **Cells**: manually annotated dataset

The Tabula Sapiens Consortium, Science (2022)

CELL TYPE ANNOTATION TOOLS (I)

Automated cell type annotation

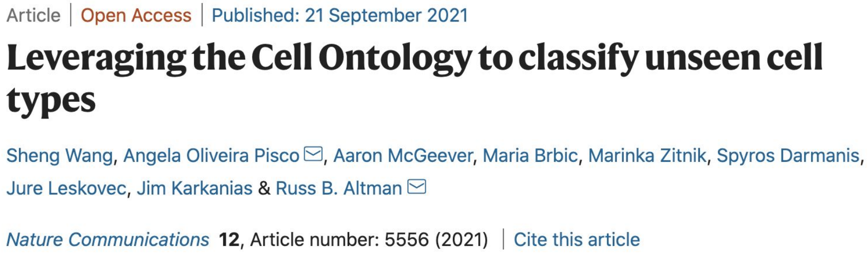


Article | Published: 19 October 2020

MARS: discovering novel cell types across heterogeneous single-cell experiments

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O. Pisco, Russ B. Altman, Spyros Darmanis & Jure Leskovec

Nature Methods (2020) | Cite this article



Article | Open Access | Published: 21 September 2021

Leveraging the Cell Ontology to classify unseen cell types

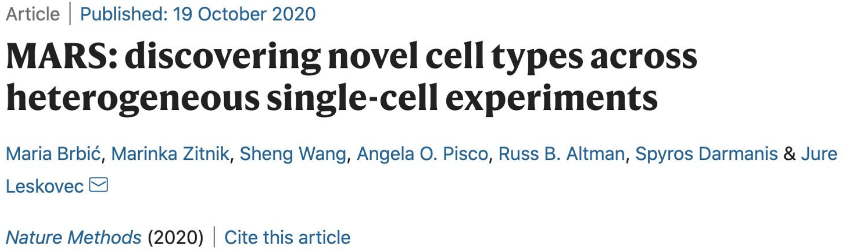
Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias & Russ B. Altman

Nature Communications **12**, Article number: 5556 (2021) | Cite this article

Brbic et al, Nat Methods (2020)
Wang, Pisco et al, Nat Comms (2021)
The Tabula Sapiens Consortium, Science (2022)

CELL TYPE ANNOTATION TOOLS (II)

Automated cell type annotation

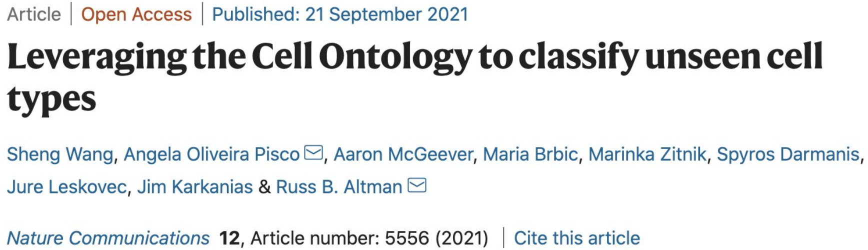


Article | Published: 19 October 2020

MARS: discovering novel cell types across heterogeneous single-cell experiments

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O. Pisco, Russ B. Altman, Spyros Darmanis & Jure Leskovec

Nature Methods (2020) | Cite this article



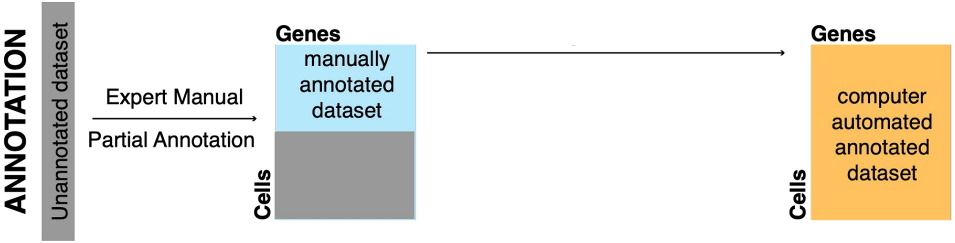
Article | Open Access | Published: 21 September 2021

Leveraging the Cell Ontology to classify unseen cell types

Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias & Russ B. Altman

Nature Communications **12**, Article number: 5556 (2021) | Cite this article

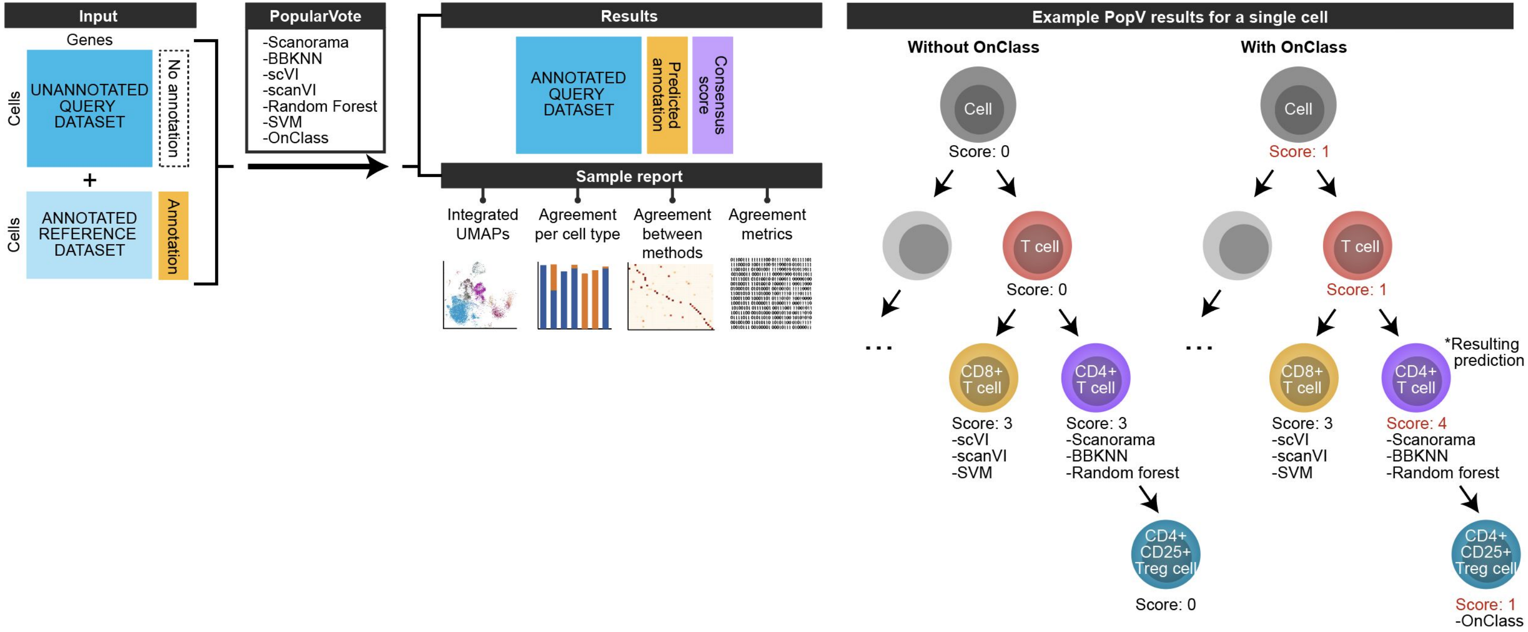
Using the Tabula datasets as general reference for annotations



Collaboration with Chenling Xu, Galen Xing, Nir Yosef

CELL TYPE ANNOTATION TOOLS (III)

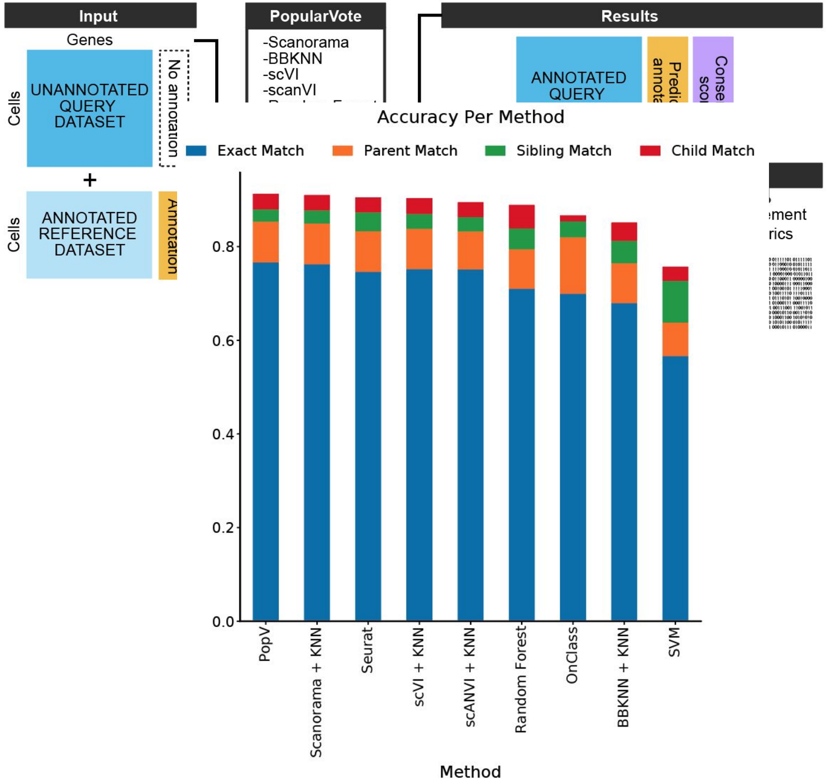
Using the Tabula datasets as general reference for annotations

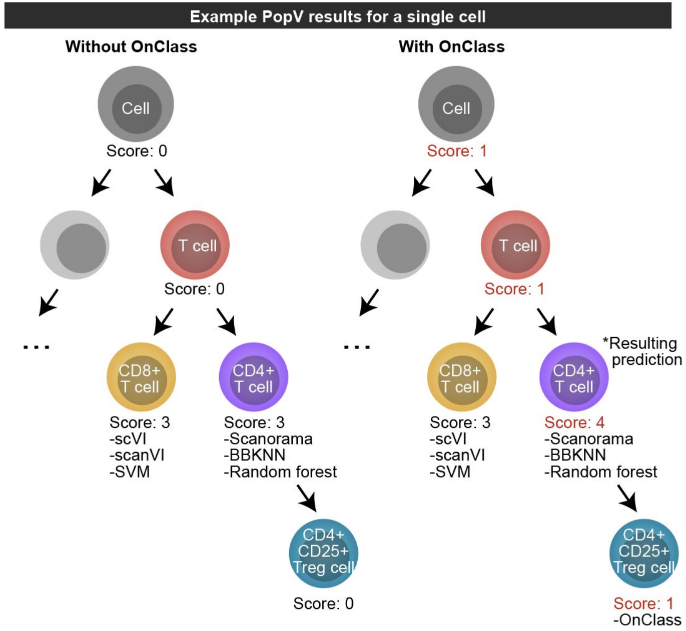


Collaboration with Chenling Xu, Galen Xing, Nir Yosef
https://tabula-sapiens-portal.ds.czbiohub.org/annotateuserdata
https://github.com/czbiohub/PopV

CELL TYPE ANNOTATION TOOLS (III)

Using the Tabula datasets as general reference for annotations

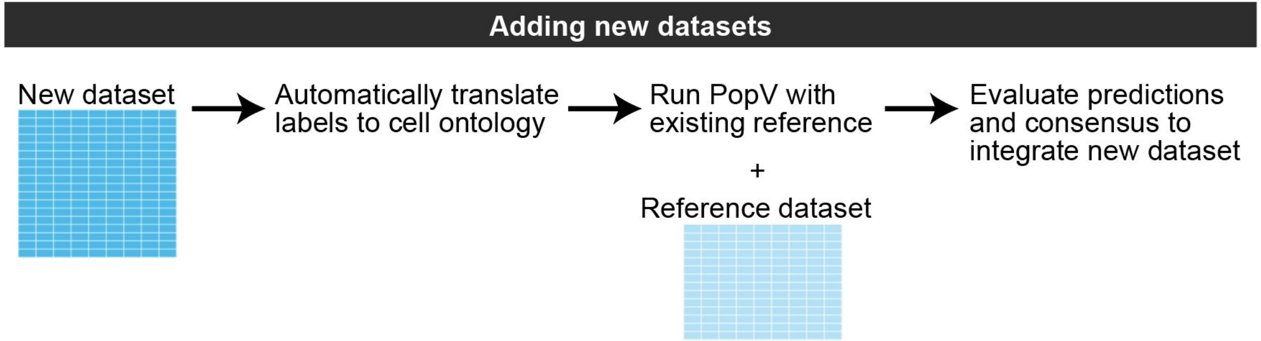




Collaboration with Chenling Xu, Galen Xing, Nir Yosef
https://tabula-sapiens-portal.ds.czbiohub.org/annotateuserdata
https://github.com/czbiohub/PopV

CELL TYPE ANNOTATION TOOLS (IV)

Using the Tabula datasets as general reference for annotations



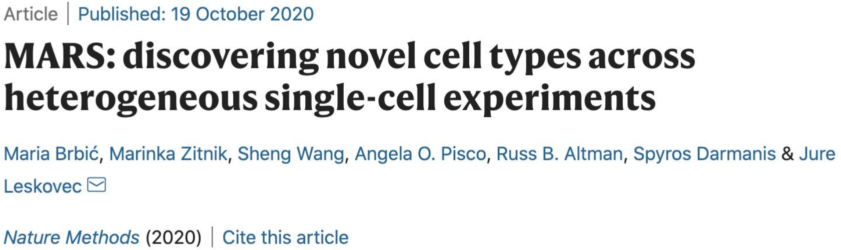
Adding new datasets

New dataset → Automatically translate labels to cell ontology → Run PopV with existing reference + Reference dataset → Evaluate predictions and consensus to integrate new dataset

Collaboration with Chenling Xu, Galen Xing, Nir Yosef
https://tabula-sapiens-portal.ds.czbiohub.org/annotateuserdata
https://github.com/czbiohub/PopV

CELL TYPE ANNOTATION TOOLS (V)

Automated cell type annotation



Article | Published: 19 October 2020

MARS: discovering novel cell types across heterogeneous single-cell experiments

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O. Pisco, Russ B. Altman, Spyros Darmanis & Jure Leskovec

Nature Methods (2020) | Cite this article

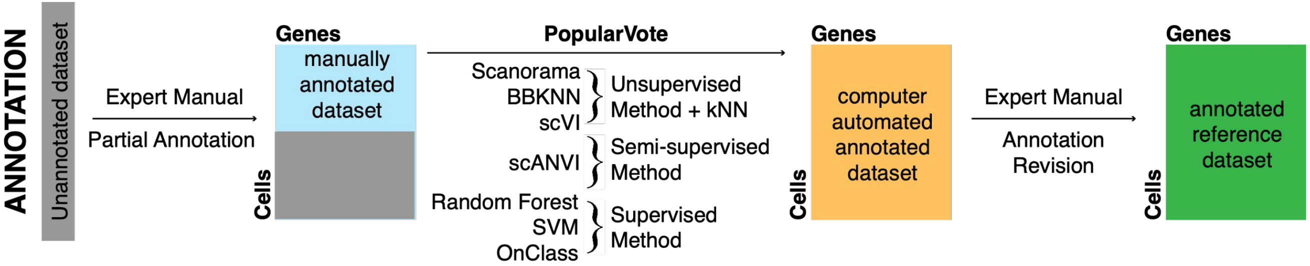
Article | Open Access | Published: 21 September 2021

Leveraging the Cell Ontology to classify unseen cell types

Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias & Russ B. Altman

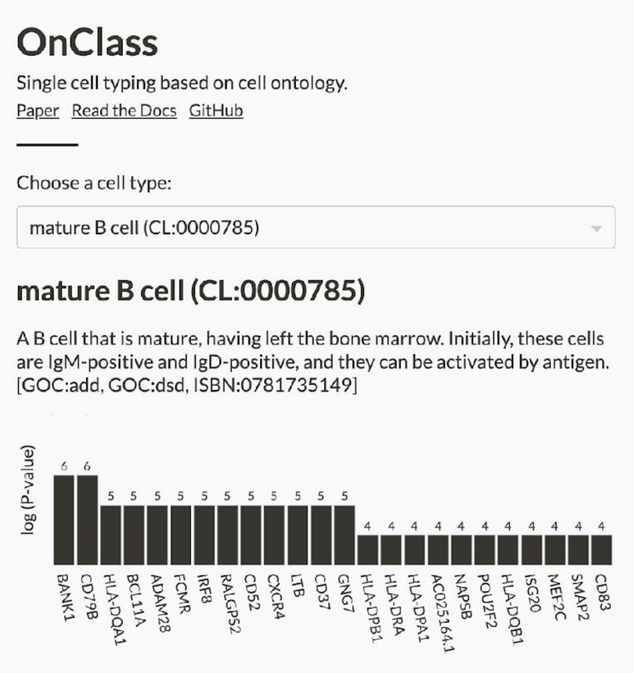
Nature Communications **12**, Article number: 5556 (2021) | Cite this article

Using the Tabula datasets as general reference for annotations



Collaboration with Chenling Xu, Galen Xing, Nir Yosef

MARKER GENES FOR CELL TYPES



OnClass

Single cell typing based on cell ontology.

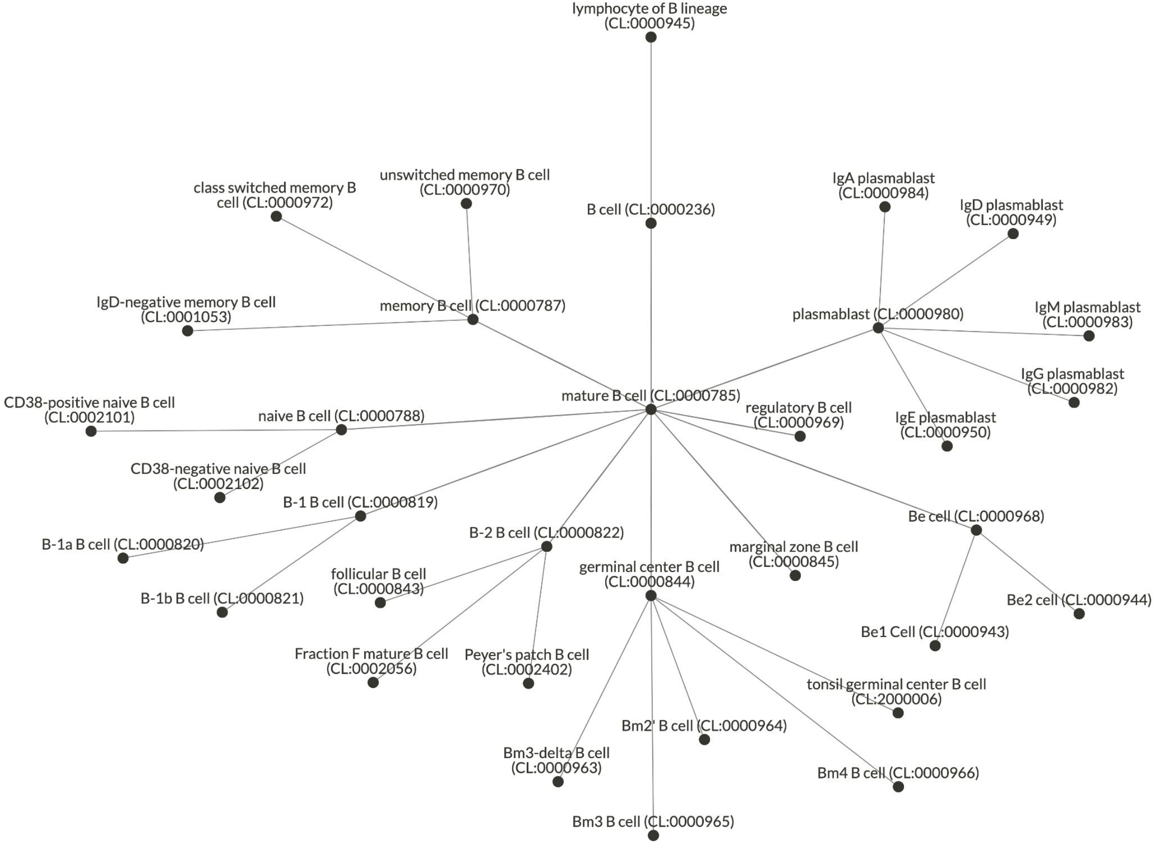
Paper Read the Docs GitHub

Choose a cell type:

mature B cell (CL:0000785)

mature B cell (CL:0000785)

A B cell that is mature, having left the bone marrow. Initially, these cells are IgM-positive and IgD-positive, and they can be activated by antigen. [GOC:add, GOC:dsd, ISBN:0781735149]



lymphocyte of B lineage (CL:0000945)
class switched memory B cell (CL:0000972)
unswitched memory B cell (CL:0000970)
B cell (CL:0000236)
IgA plasmablast (CL:0000984)
IgD plasmablast (CL:0000949)
IgD-negative memory B cell (CL:0001053)
memory B cell (CL:0000787)
plasmablast (CL:0000980)
IgM plasmablast (CL:0000983)
IgG plasmablast (CL:0000982)
mature B cell (CL:0000785)
CD38-positive naive B cell (CL:0002101)
naive B cell (CL:0000788)
regulatory B cell (CL:0000969)
IgE plasmablast (CL:0000950)
CD38-negative naive B cell (CL:0002102)
B-1 B cell (CL:0000819)
Be cell (CL:0000968)
B-1a B cell (CL:0000820)
B-2 B cell (CL:0000822)
marginal zone B cell (CL:0000845)
germinal center B cell (CL:0000844)
follicular B cell (CL:0000843)
B-1b B cell (CL:0000821)
Be2 cell (CL:0000944)
Be1 Cell (CL:0000943)
Fraction F mature B cell (CL:0002056)
Peyer's patch B cell (CL:0002402)
tonsil germinal center B cell (CL:2000006)
Bm2' B cell (CL:0000964)
Bm3-delta B cell (CL:0000963)
Bm4 B cell (CL:0000966)
Bm3 B cell (CL:0000965)

Wang, Pisco, et al, Nat Comms (2021)
The Tabula Sapiens Consortium, Science (2022)

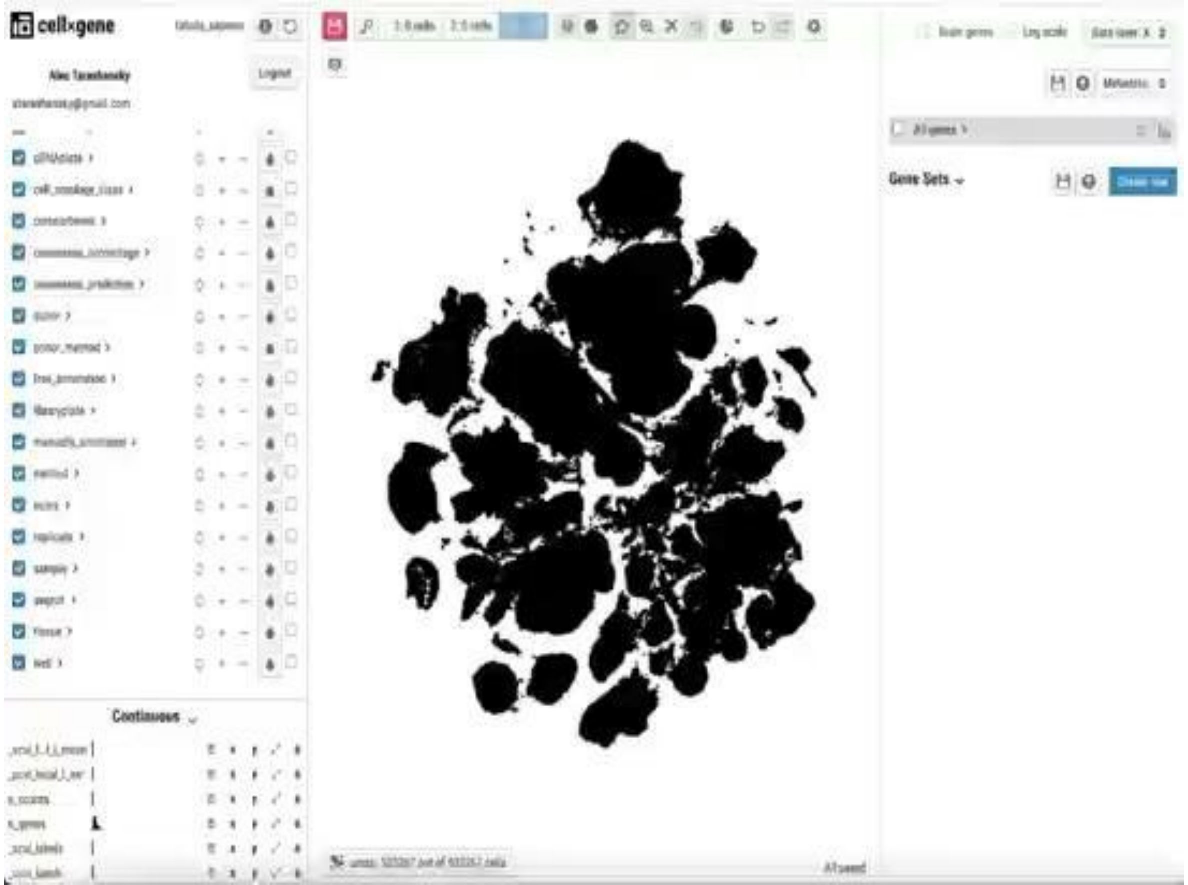
EXCELLXGENE

cellxgene

Exploratory CellxGene (ExCellxGene)

- Lidar
- Differential gene expression on the fly
- Nested embeddings
- Leiden clustering
- Re-embeddding
- Sankey diagrams

https://github.com/czbiohub/excellxgene



THANK YOU!

The Tabula Muris consortium

The Tabula Sapiens consortium

The Tabula Microcebus consortium

The Fly Cell Atlas consortium

The Covid Tissue Atlas consortium

CZBiohub Data Science Team

CZBiohub Genomics Team

CHAN ZUCKERBERG BIOHUB

We are hiring!

Cal

UCSF
University of California
San Francisco

Stanford
University

DONOR NETWORK WEST

Ahmad Salehi
Ravi Ponnusmi

We express our gratitude and thanks to donor WEM and his family, as well as all of the anonymous organ and tissue donors and their families for giving both the gift of life and the gift of knowledge by their generous donations.

cell×gene

CZI
Collin Megill
Max Lombardo
Ambrose Carr
Jenn Tang
Tiago Carvalho

Fabian Theis

Fabian Theis, Helmholtz Munich, Germany

Learning and using gene set ontologies in single cell genomics

Data management & annotation with sfaira

Data curation:

- one loader per study, publicly maintained
- easy to contribute, easy to extend
- allow generation of streamlined .h5ad files, e.g: *cellxgene*

Data collection management:

- select datasets from a collection based on meta data
- interface local and public (e.g. cellxgene) data collections
- prepare data for model training

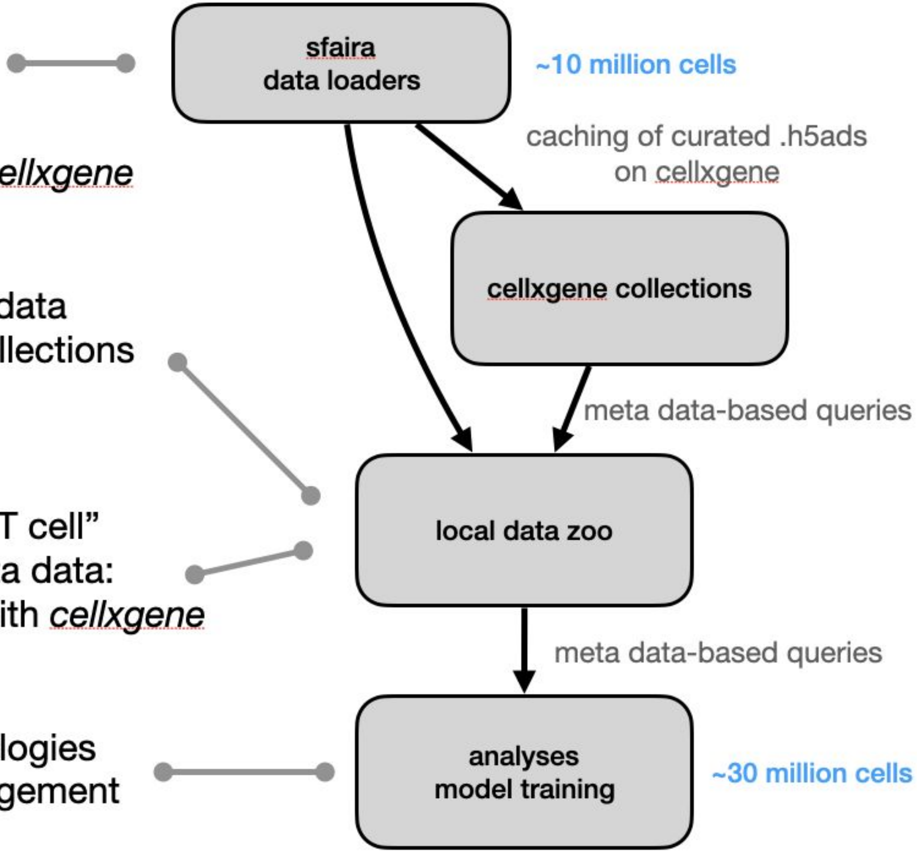
Leverage meta data:

- ontology-based queries to data: subset to "is a T cell"
- uses publicly curated ontolgies for all major meta data: cell type, tissue, organism, disease, …, synced with *cellxgene*

Models:

- syntax for feature and label space based on ontologies

—> seamless integration with data collection management



sfaira data loaders

~10 million cells

caching of curated .h5ads on cellxgene

cellxgene collections

meta data-based queries

local data zoo

meta data-based queries

analyses model training

~30 million cells

Case study: Preparing HLCA for cellxgene

Data curation:

- map meta data annotation to ontologies
- format fields in AnnData object to satisfy cellxgene requirements

effective time for writing data loader: ~ 1-2 hours

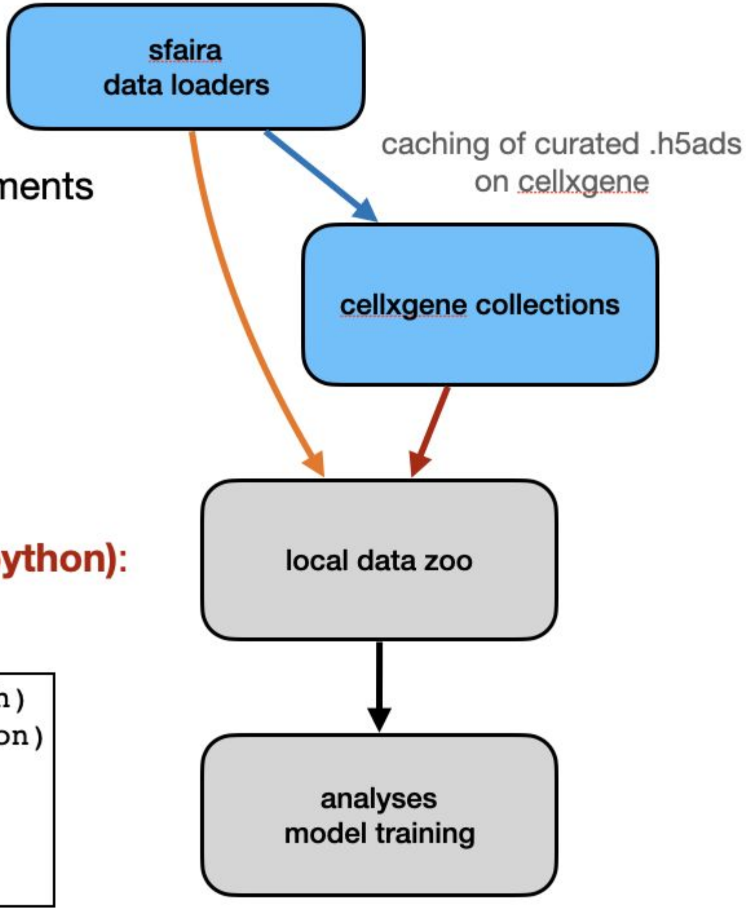
Data collection management:

Load HLCA AnnData via either mechanism:

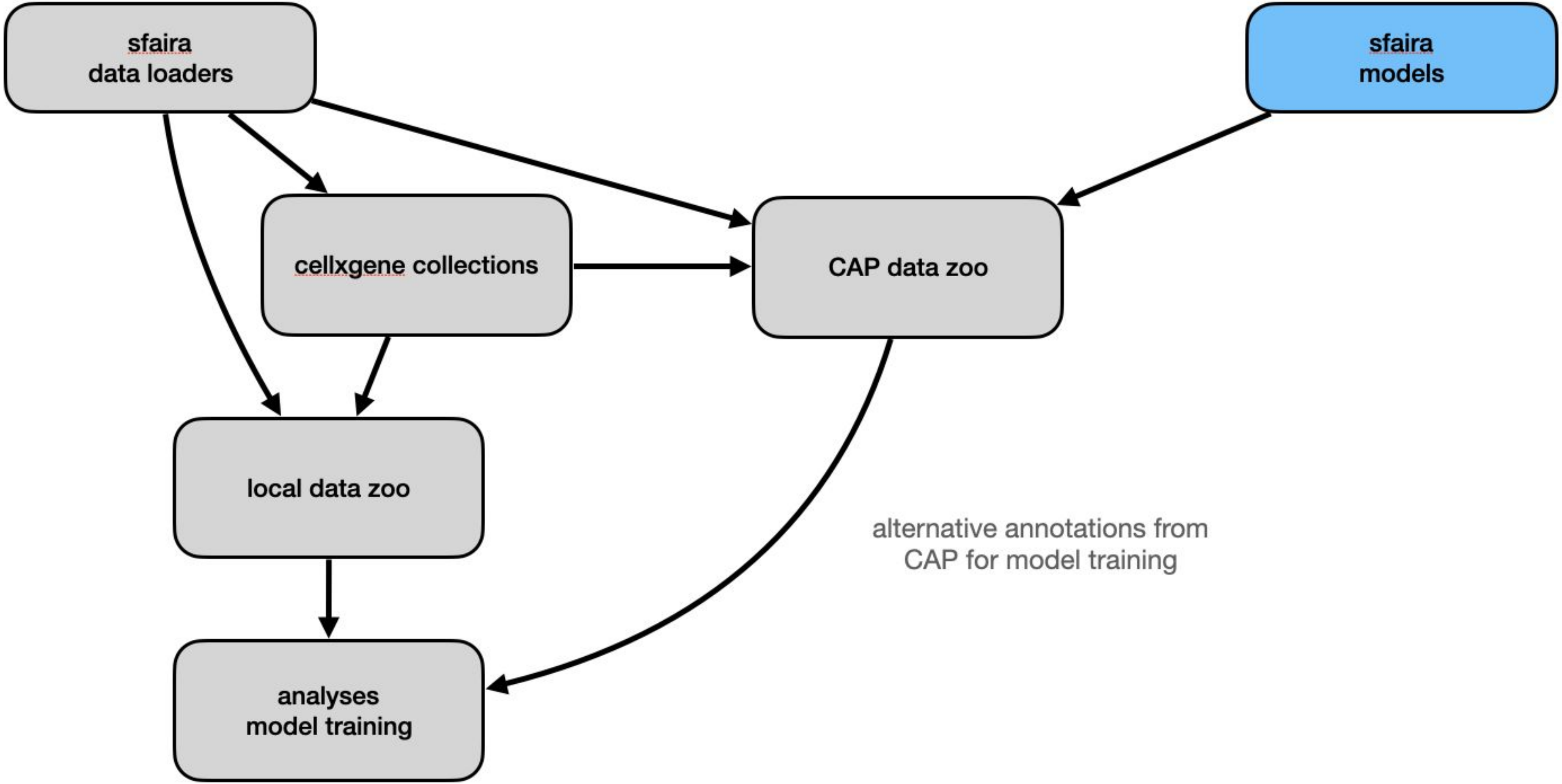
- via sfaira data loader:
  - load non-curated (as published in paper) AnnData
  - load curated AnnData (any curation format)
- via cellxgene **(website)** or sfaira interface of cellxgene **(in python)**:
  - load curated AnnData (cellxgene format)

sfaira interface of cellxgene

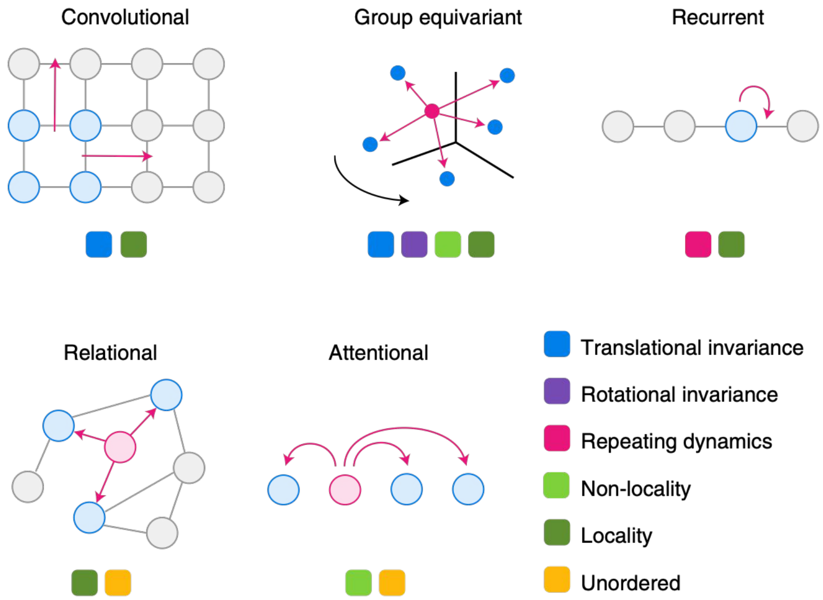
```
dsg = DatasetSuperGroupDatabases(data_path=cache_path)
dsg.subset(key="collection_id", values=HLCA_collection)
dsg.download()
dsg.load()
adatas_hlca = dsg.adata_ls
```



Sfaira, CAP, cellxgene



Outlook: differential biology - deep learning for modeling molecular mechanisms



Convolutional

Group equivariant

Recurrent

Relational

Attentional

🟦 Translational invariance

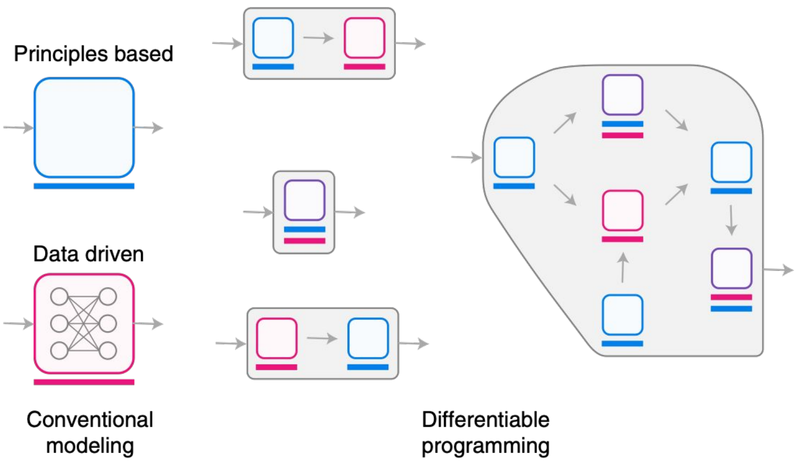
🟪 Rotational invariance

🟥 Repeating dynamics

🟩 Non-locality

🟩 Locality

🟧 Unordered



Principles based

Data driven

Conventional modeling

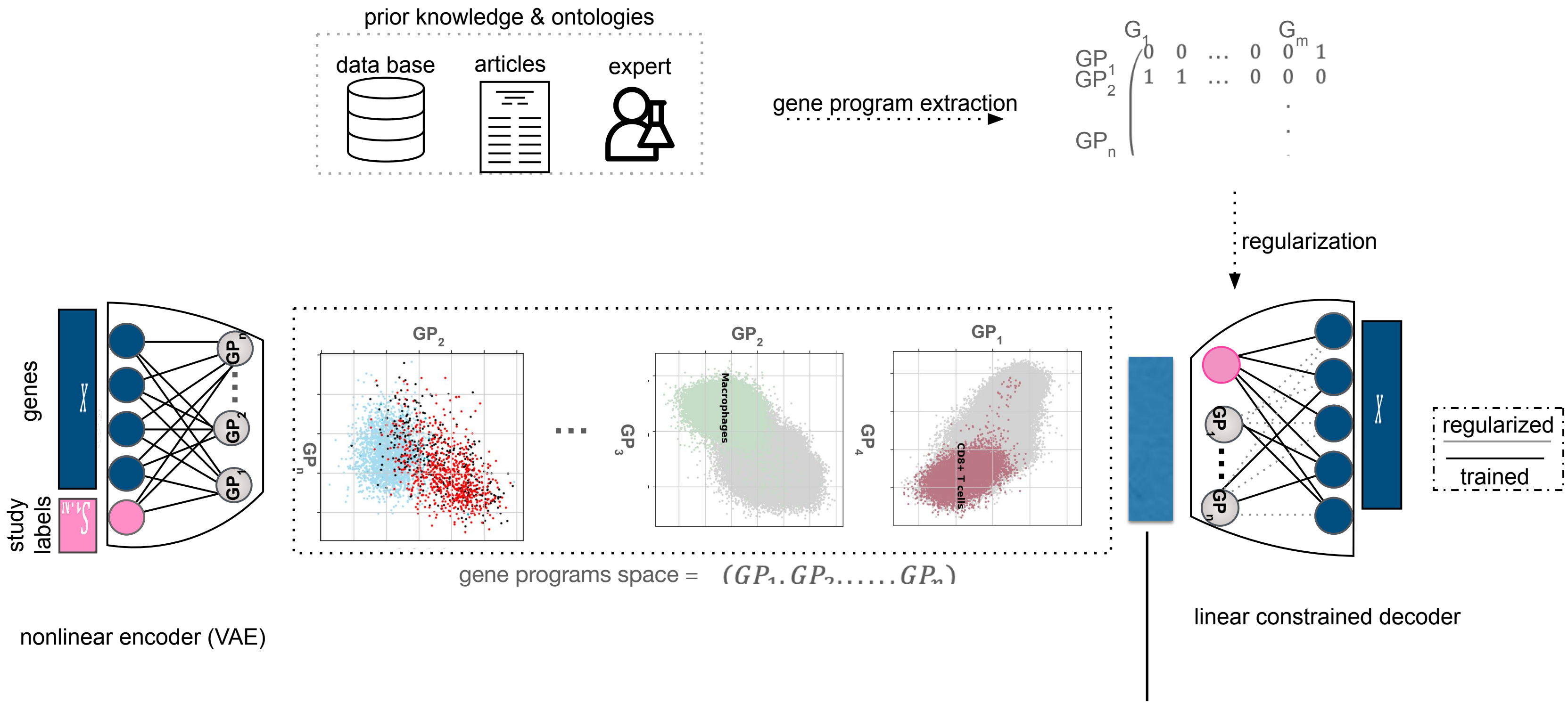
Differentiable programming

network primitives and invariances

→ questions: reuse primitives? add constraints?

AlQuraishi & Sorger, *Nat Methods* 2021

Interpretable atlas querying: explainable programmable mapper (expiMap)



prior knowledge & ontologies

data base

articles

expert

gene program extraction

regularization

gene programs space = $(GP_1, GP_2, ..., GP_n)$

nonlinear encoder (VAE)

linear constrained decoder

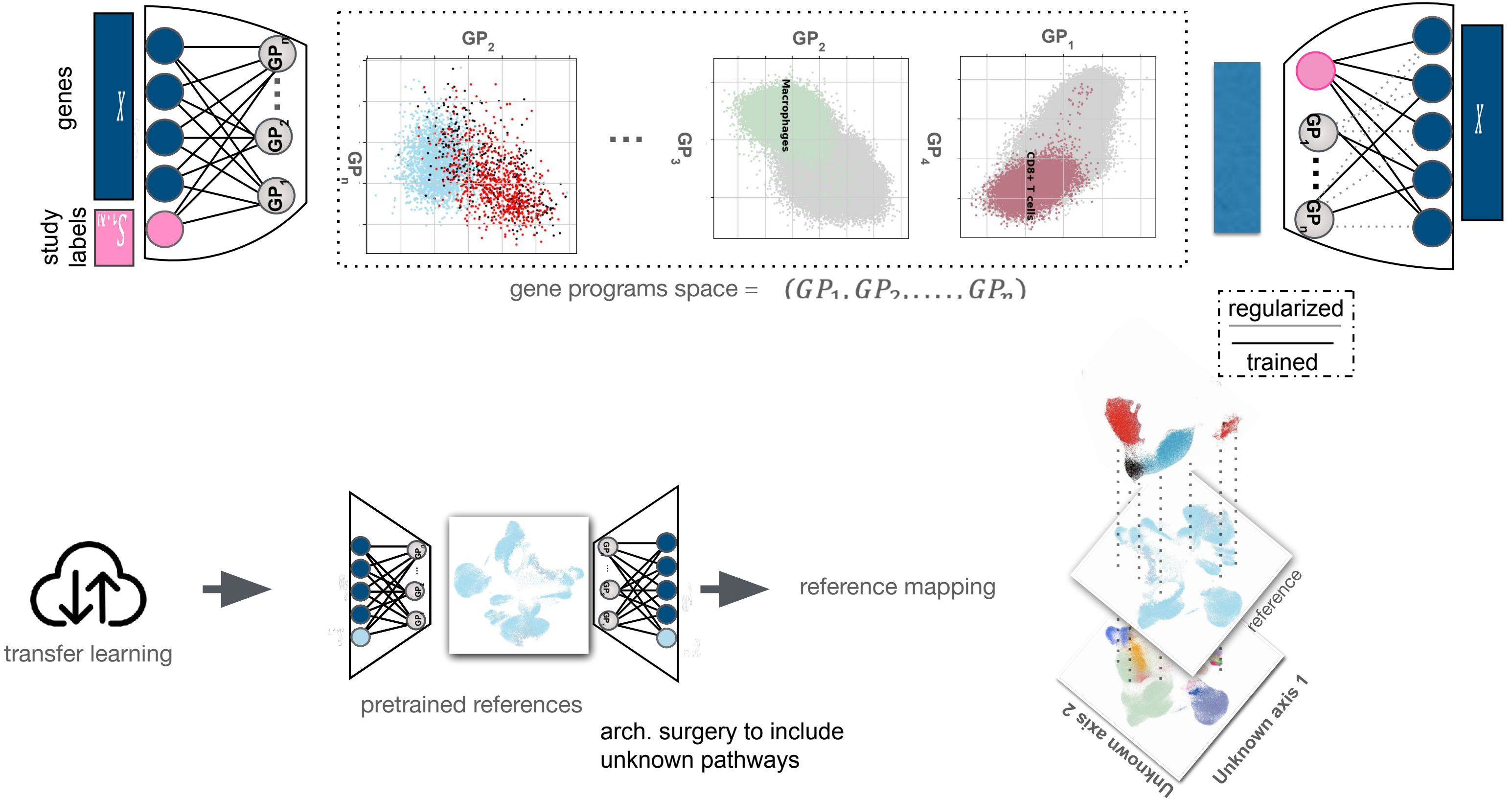
regularized

trained

Sergei Rybakov, Mo Lotfollahi

Lotfollahi et al biorxiv 2022

Interpretable atlas querying: explainable programmable mapper (expiMap)



gene programs space = $(GP_1, GP_2, \ldots\ldots, GP_n)$

transfer learning

pretrained references

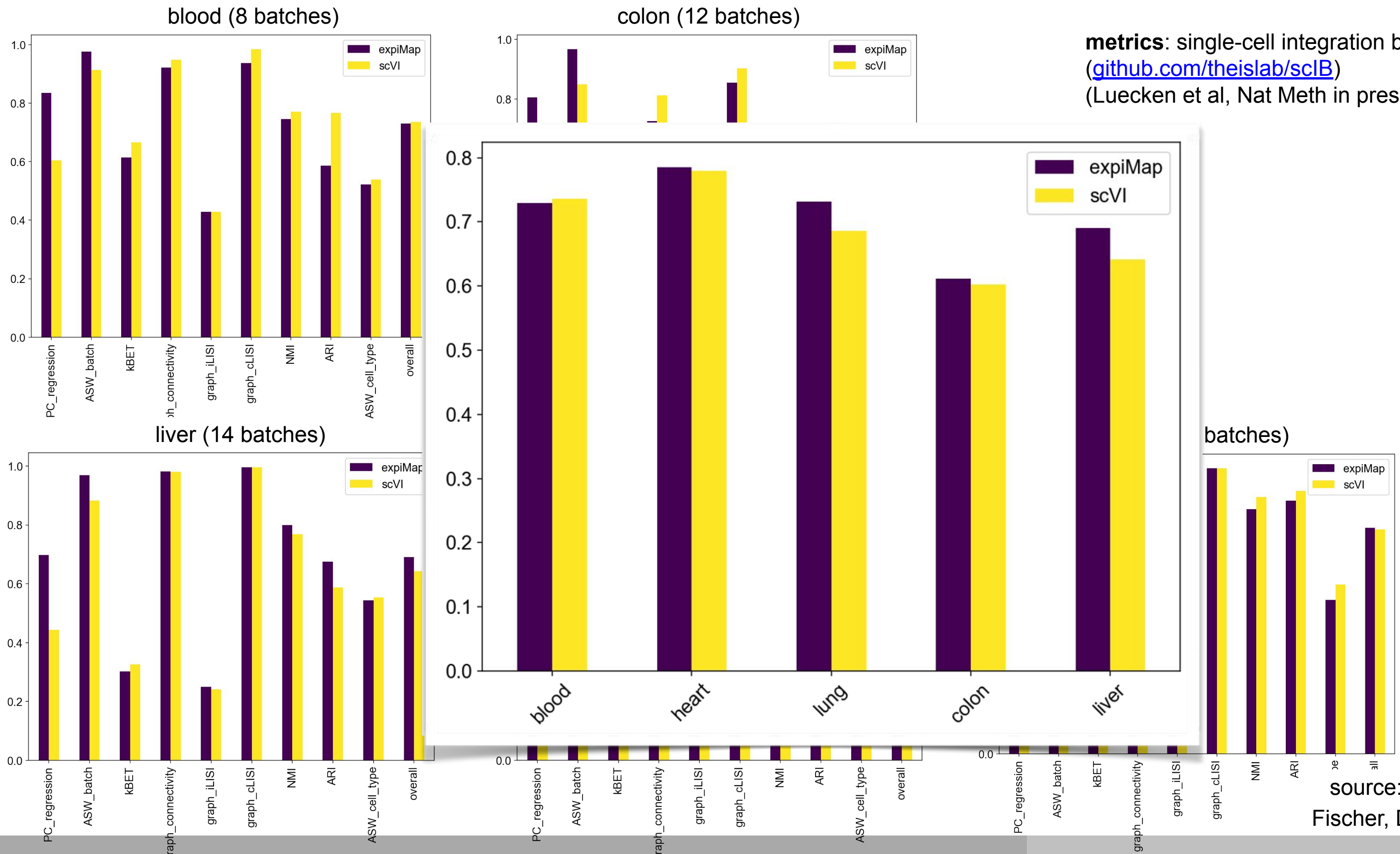
arch. surgery to include unknown pathways

reference mapping

Sergei Rybakov, Mo Lotfollahi

Lotfollahi et al biorxiv 2022

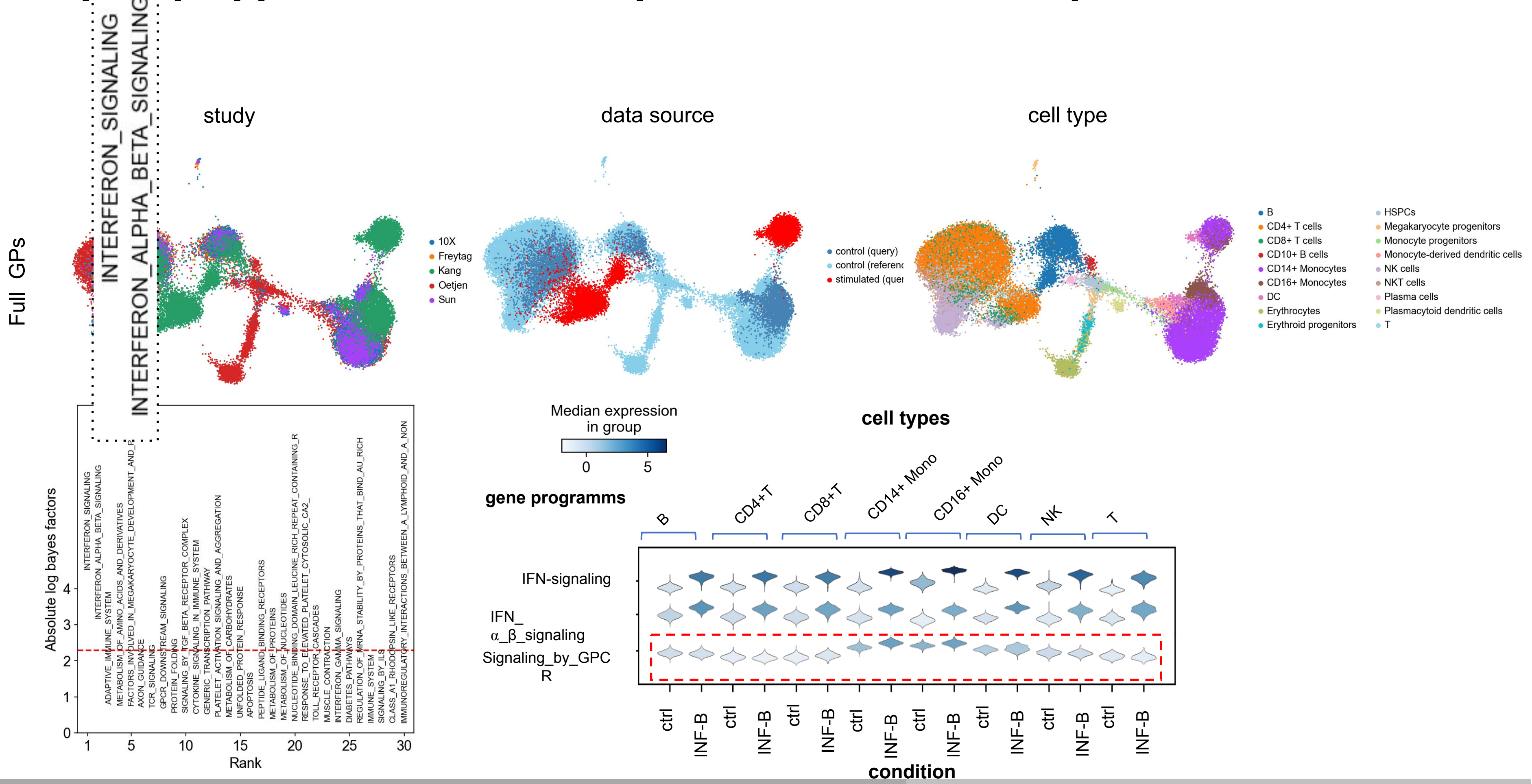
expiMap does not loose expressiveness versus nonlinear models



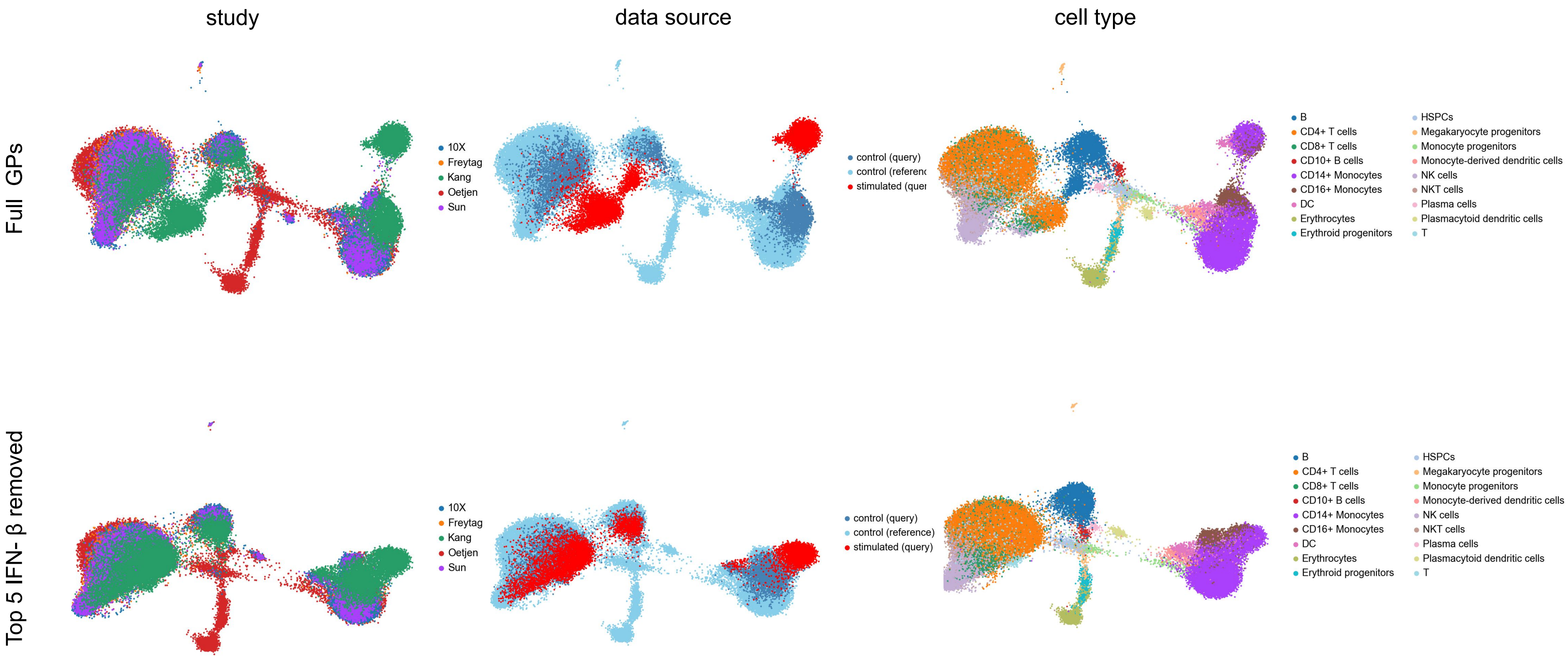
metrics: single-cell integration benchmark (github.com/theislab/scIB) (Luecken et al, Nat Meth in press)

source: sfaira multi-study organ batches
Fischer, Dony et al, *Genome Biology* 2021

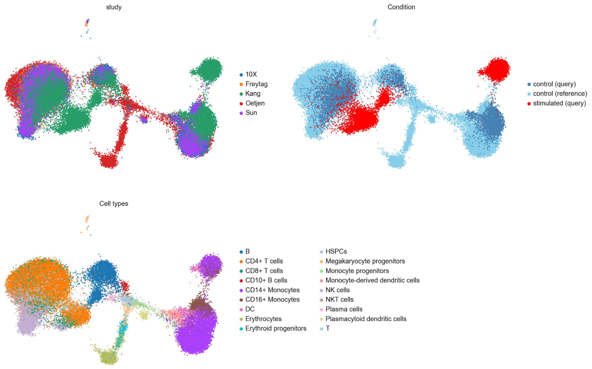
exp ap approximates latent space structure in interpretable fashion



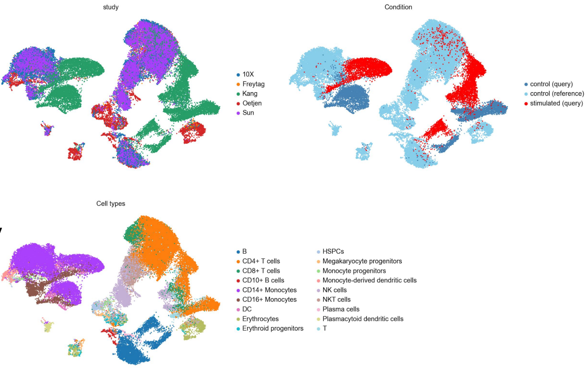
expiMap approximates latent space structure in interpretable fashion



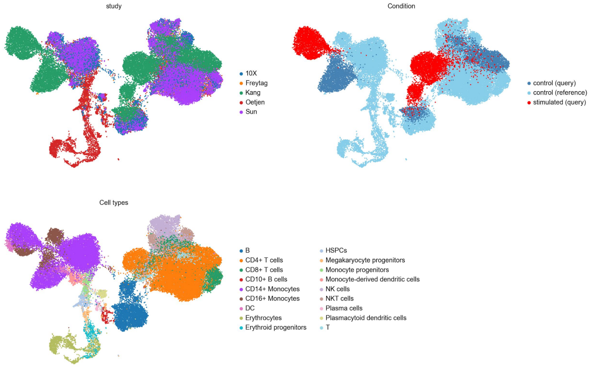
expiMap integrates cross experiment data while retaining perturbation effect



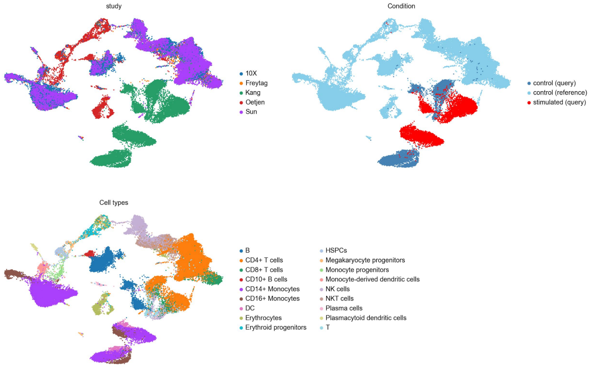
expiMa
p



Symphon
y



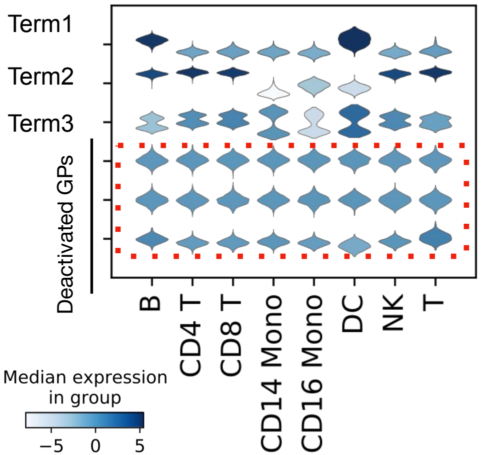
scVI

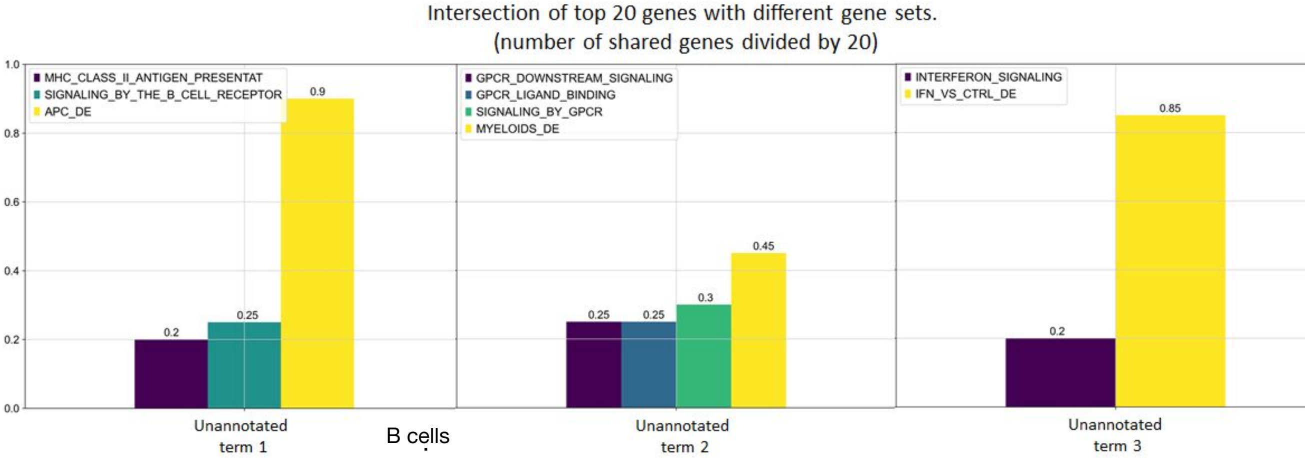


Seurat

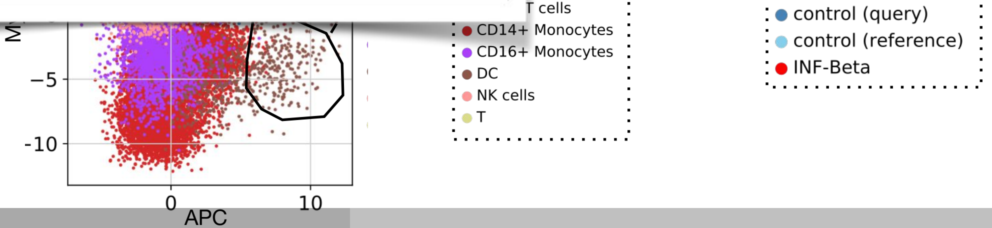
Learning new interpretable programs

removed B cells & related pathways (incl Inf) from reference
-> challenge model to find them during query mapping




Intersection of top 20 genes with different gene sets. (number of shared genes divided by 20)

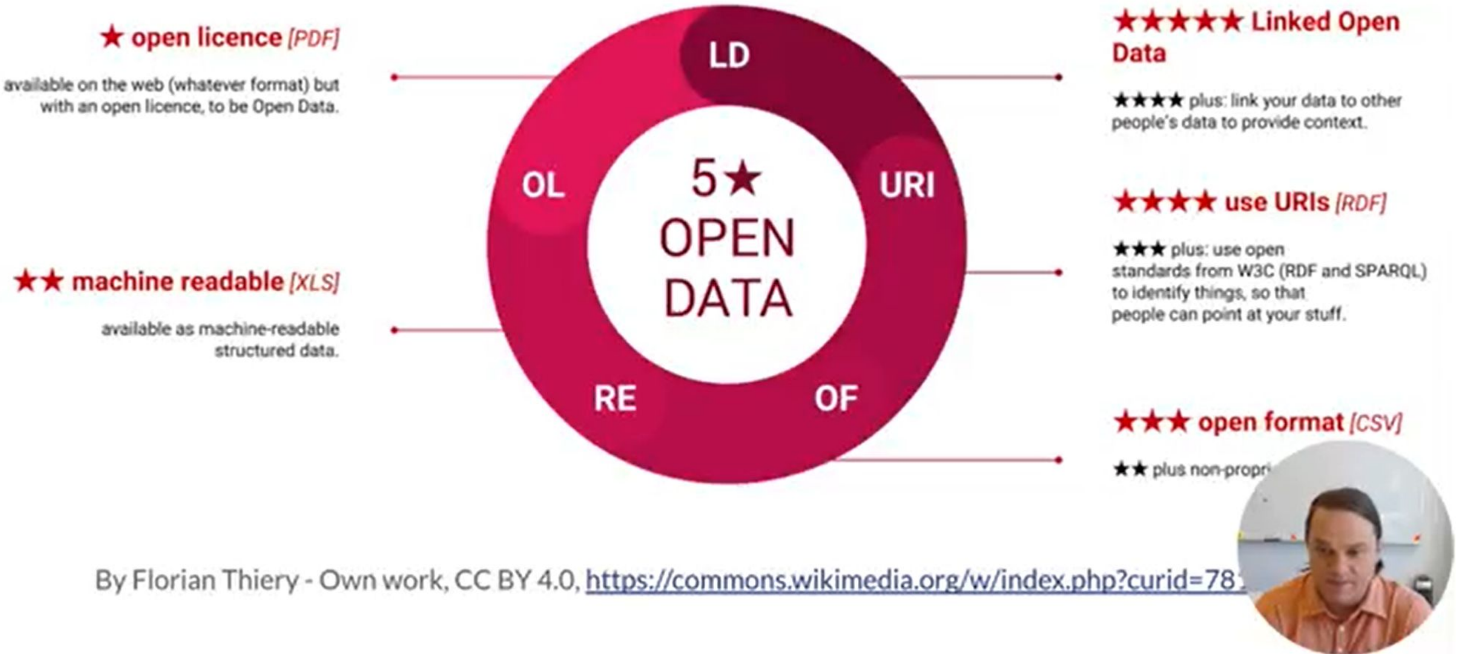
ongoing: transferring learned interpretable embedding from PBMCs to Covid samples helps identifying differential communication pathway during moderate and severe COVID-19



Bruce W. Herr II

Bruce W. Herr II, Indiana University, USA

Slides | Video



★ open licence *[PDF]*

available on the web (whatever format) but with an open licence, to be Open Data.

★★ machine readable *[XLS]*

available as machine-readable structured data.

★★★★★ Linked Open Data

★★★★ plus: link your data to other people's data to provide context.

★★★★ use URIs *[RDF]*

★★★ plus: use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.

★★★ open format *[CSV]*

★★ plus non-propri

By Florian Thiery - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=78

Discussion

Discussion Topics

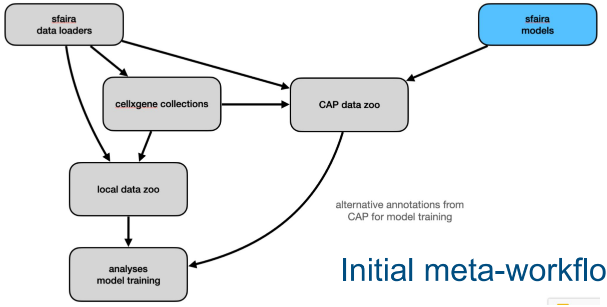
1. How can we **extend and improve ontologies** as our knowledge grows leveraging expert input, experimental data and feedback from different atlasing efforts?
2. How can we **make ontology annotation, following standard schemas, easy**, efficient, and accurate while leaving room for revising and adding to existing ontologies?
3. How can we enable downstream users to take advantage of ontology structure and content in **analysis, visualization and machine learning pipelines/applications**?
4. How can improved annotation with ontologies and the use of linked open data (LOD) help us to interlink atlas data and from multiple consortia and **construct more integrated, coherent, and queryable atlases**?

Report Back

Cell Ontology, Annotation & Metadata Breakout - Report

Next Steps

- **Agree on Portal Workflow(s):** Where should users go to annotate their data?
  - How should data flow between portals?
  - Should we encourage annotation to integrated/consensus/cross modal analysis?
- **Agree on** Versioning for ATLAS data and portals
  - So that we can track change as data moves.
- **Agree on** Dataflows for new cell type claims
  - Provisional Cell Ontology (semi-automated)
  - Cell Ontology (curated)



sfaira data loaders

sfaira models

cellxgene collections

CAP data zoo

local data zoo

analyses model training

alternative annotations from CAP for model training

Initial meta-workflow by Fabian Theis

Other Recommendations

- Retain user free text annotations in addition to ontology annotations
- Include confidence scores with cell type projections
- Cell annotation: Evidence for cell types is needed - but how can we record it well
  - Markers? Projection algorithms? Free text?
- Integrating with spatial data
  - We need reference atlases for anatomical regions.
  - Cell segmentation: Collect gold standards for anatomical and cell segmentations. Run algorithm comparisons, e.g., via Kaggle (HuBMAP+HPA "Hacking the Human Body")

https://cns-iu.github.io/workshops/2022-06-27_human_cell_atlas