Carnegie Mellon University

wellcome sanger institute

UC SANTA CRUZ

HIVE-TC CMU: Comprehensive, Flexible and FAIR Tools for the HuBMAP HIVE





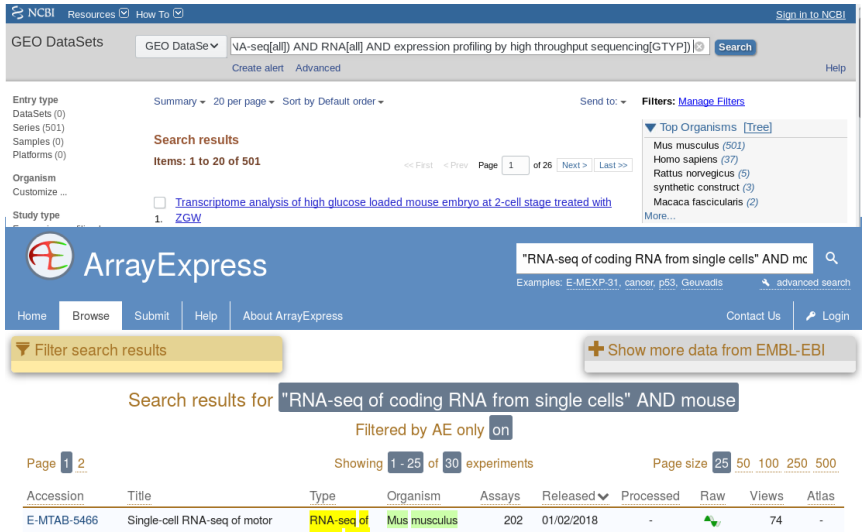


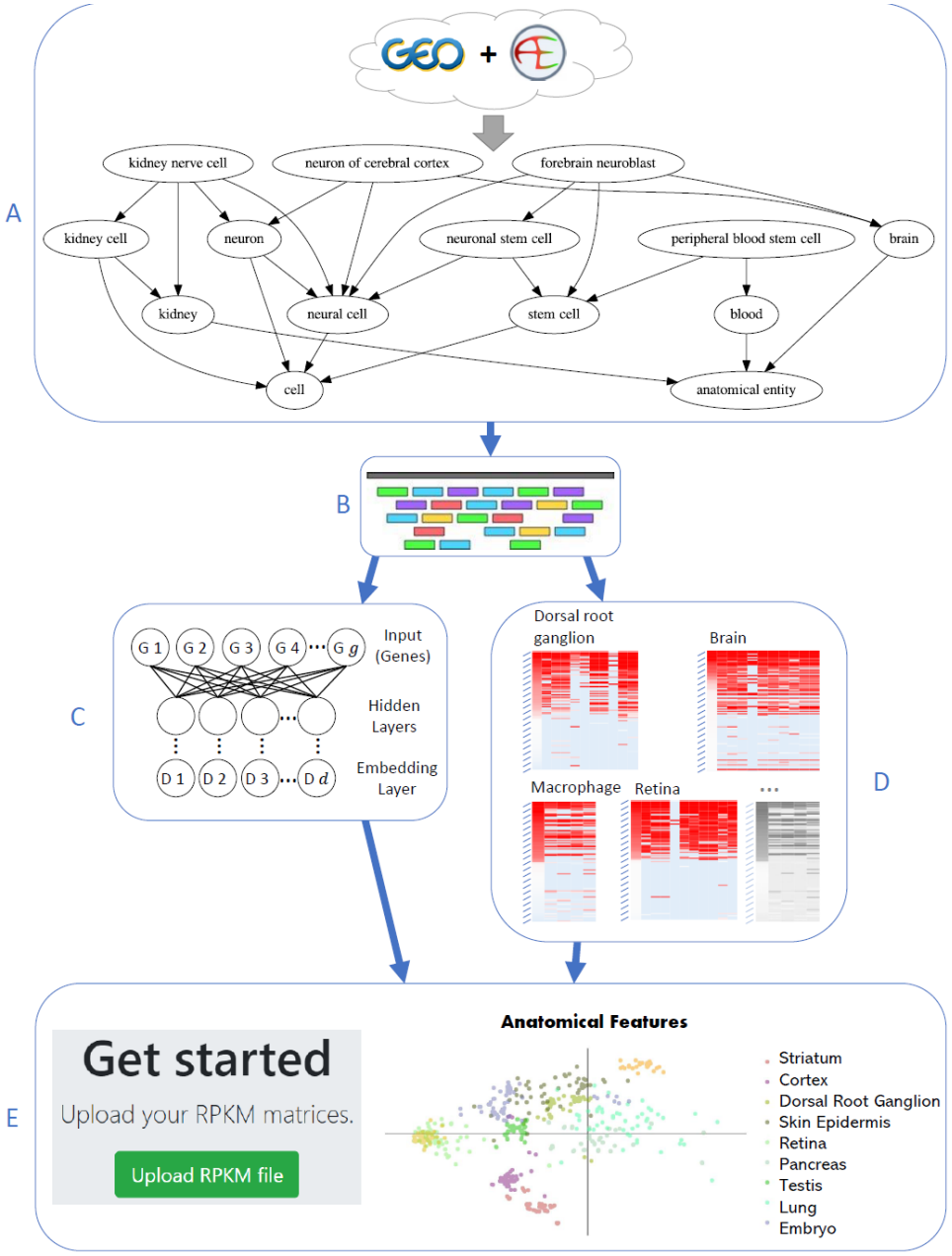
2019 HuBMAP Meeting

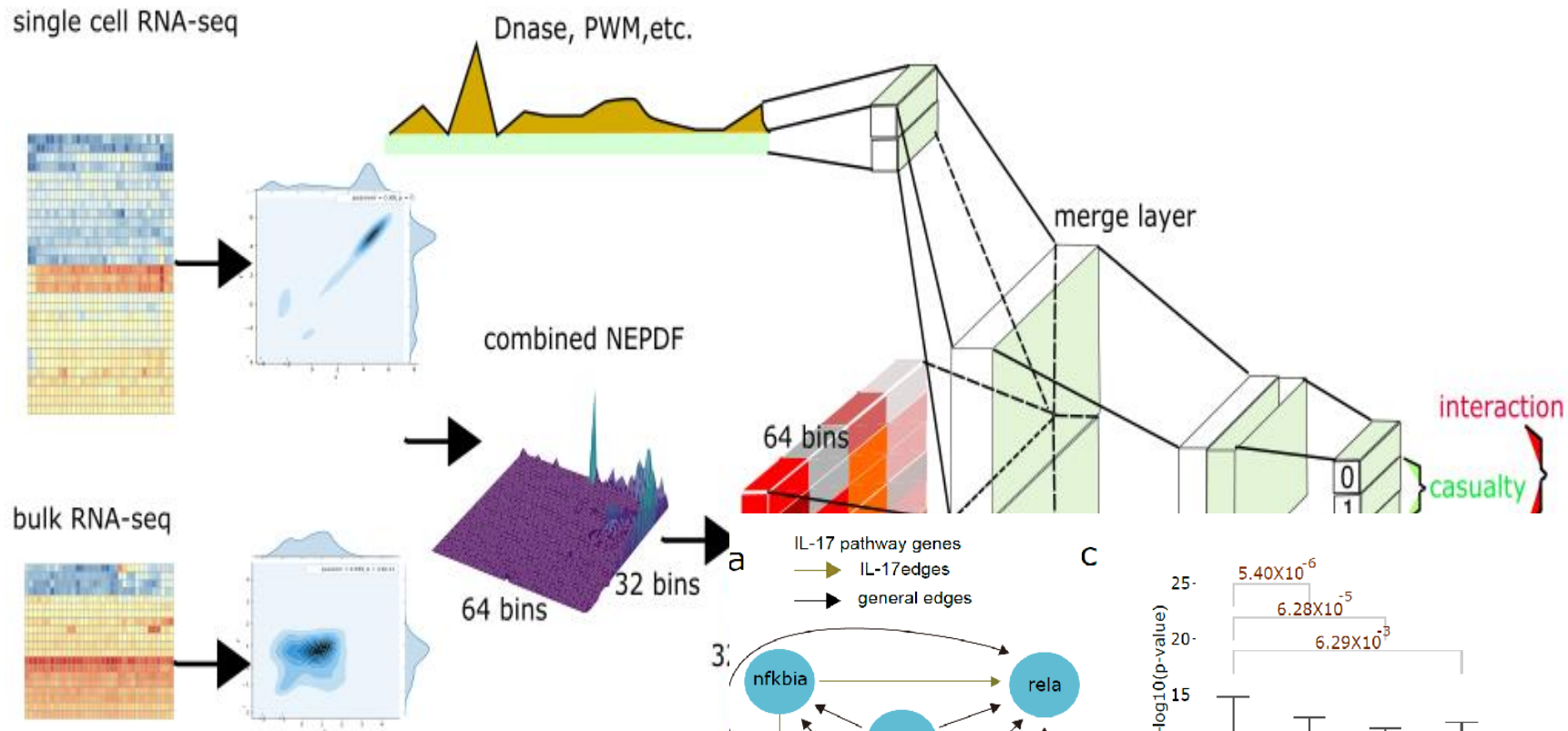
Major focus in Year 1

- Our major focus in the first year has been on:
  - Pipelines for processing HuBMAP data
  - Working on the data portal and website
  - Connecting with other efforts (mainly HCA)
- Most important issues we have addressed so far:
  - Establishing and working with the DRT's to fully define the set of pipeline needed for the initial release
  - Initial implementation and testing of containerized pipelines for sequencing and imaging data
  - Imaging data workshop
  - Minimizing redundancy with work at HCA
  - Initial development and discussions about the HuBMAP portal

Automatically collecting and processing scRNA-Seq data

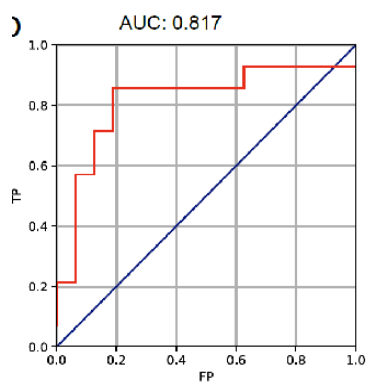


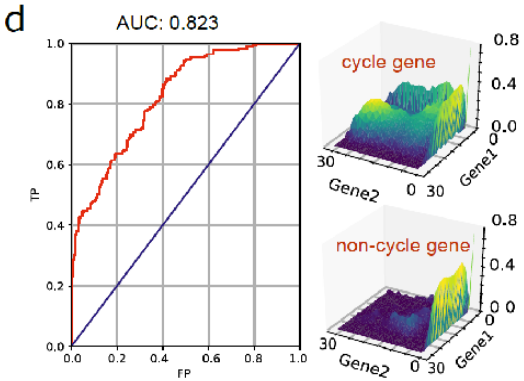


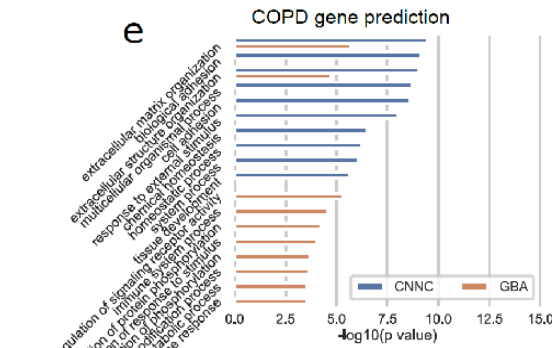

single cell RNA-seq
Dnase, PWM,etc.
merge layer
combined NEPDF
64 bins
32 bins
interaction
casualty
bulk RNA-seq

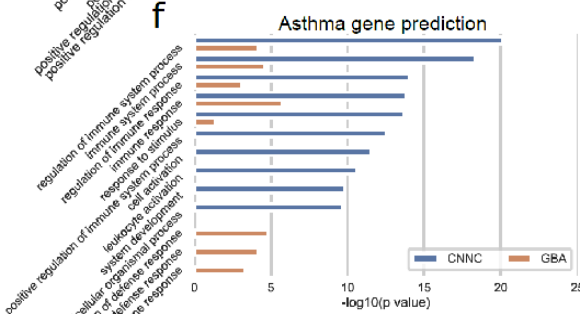
a
IL-17 pathway genes
IL-17edges
general edges
nfkbia
rela
ikbkb
nfkb1
traf6
map3k7

c
5.40X10-6
6.28X10-5
6.29X10-3
-log10(p-value)
CNNC
SC&bulk PC
SC PC
Bulk PC
method


AUC: 0.817
TP
FP


d
AUC: 0.823
TP
FP
cycle gene
non-cycle gene
Gene1
Gene2


e
COPD gene prediction
CNNC
GBA
-log10(p value)


f
Asthma gene prediction
CNNC
GBA
-log10(p value)

Sequencing Pipelines

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Priority	Assay		Institute	Tissue	Responsible	How do you get from bcl to usable data?		Steps in blue are performed at the HIVE (light blue 3)	Steps in red are performed at TMCs, but results should be uploaded to the HIVE (light			
2		Genomics (SNP)	custom SNP Chip (~924K SNPs)	U of FL	spleen	Maigan							
3	1	Genomics)	expression profile of individual	U of FL	Lymph Nodes	Maigan	(bcl2fastq + 10x	(align (STAR)/filter/					
4		sci-RNAseq	expression profile of individual	UW	(9 anatomic sites)	Hannah	bcl2fastq	custom dmux	Trim with trim_galore	genes plus 500bp	custom script	by cell matrix	
5		snRNAseq	provides the expression profile of	UCSD	trachea, bronchus,	Blue	(bcl2fastq)	using deindexer	sample (Custom	generation of digital	merging - R		
6	1	scRNAseq	expression profile of individual	Stanford	sites), Large Bowel	Hilton							
7	2	bulk RNAseq	expression profile of a sample	Stanford	sites), Large Bowel	Hilton							
8	3	bulk ATACseq	Assay for Transposase-Accessible Chromatin using sequencing is a technique used to study chromatin accessibility.	Stanford	Small Bowel (4 sites), Large Bowel (4 sites), Small Bowel jejunum, Left Colon	Aaron H/Jason Hilton							
9	2	bulk RNAseq	Bulk RNA-seq (In house protocol)	UCSD	trachea, bronchus,	Blue							
10		sci-ATACseq	Accessible Chromatin using sequencing is a technique used to study chromatin accessibility.	UW	Endothelium, Heart (9 anatomic sites)	Hannah	bcl2fastq	barcode correction with custom script	Trim with trim_galore	Align with bowtie2	dedup with custom script	Call peaks with MACS2	by cell mat with custor script
11		Accessability seq	accessibility profiles in individual	UCSD	trachea, bronchus,	Blue	(bcl2fastq)	using deindexer	sample (Custom	by-bin matrix			
12		scATACseq	Single Cell Assay for Transposase-Accessible Chromatin using sequencing is a technique used to study chromatin accessibility.	Stanford	Small Bowel (4 sites), Large Bowel (4 sites), Small Bowel jejunum, Left Colon								
13		sequencing	Measure DNA methylation										
14													
15													
16													
17													
18													
19													
20													
21													

Sheet1

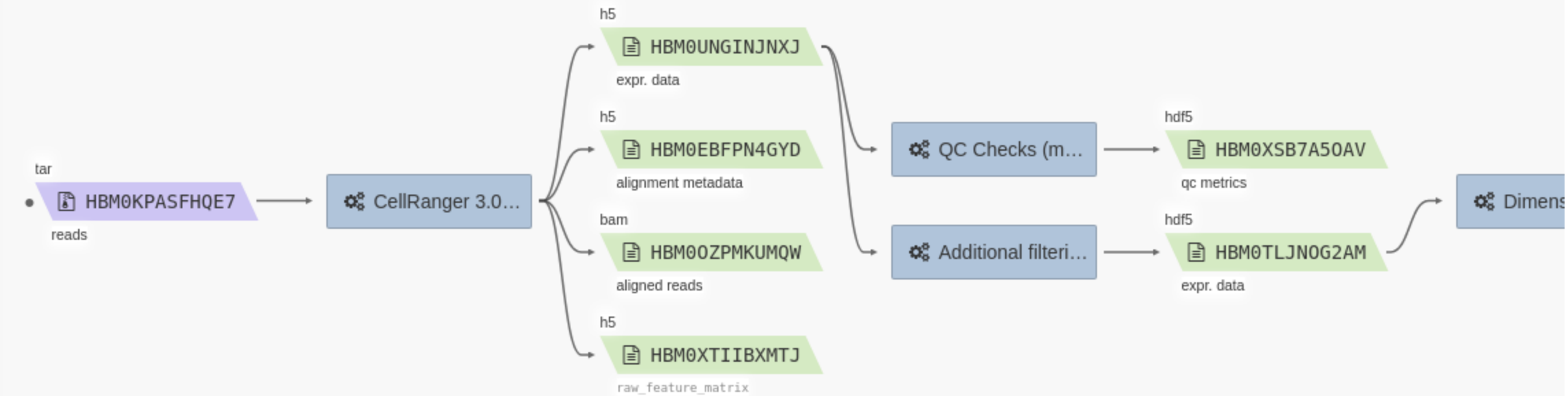
Graph & Summary | Details | 0 Audit

Graph

Show Reference Files | Show Parameters

Analysis Steps

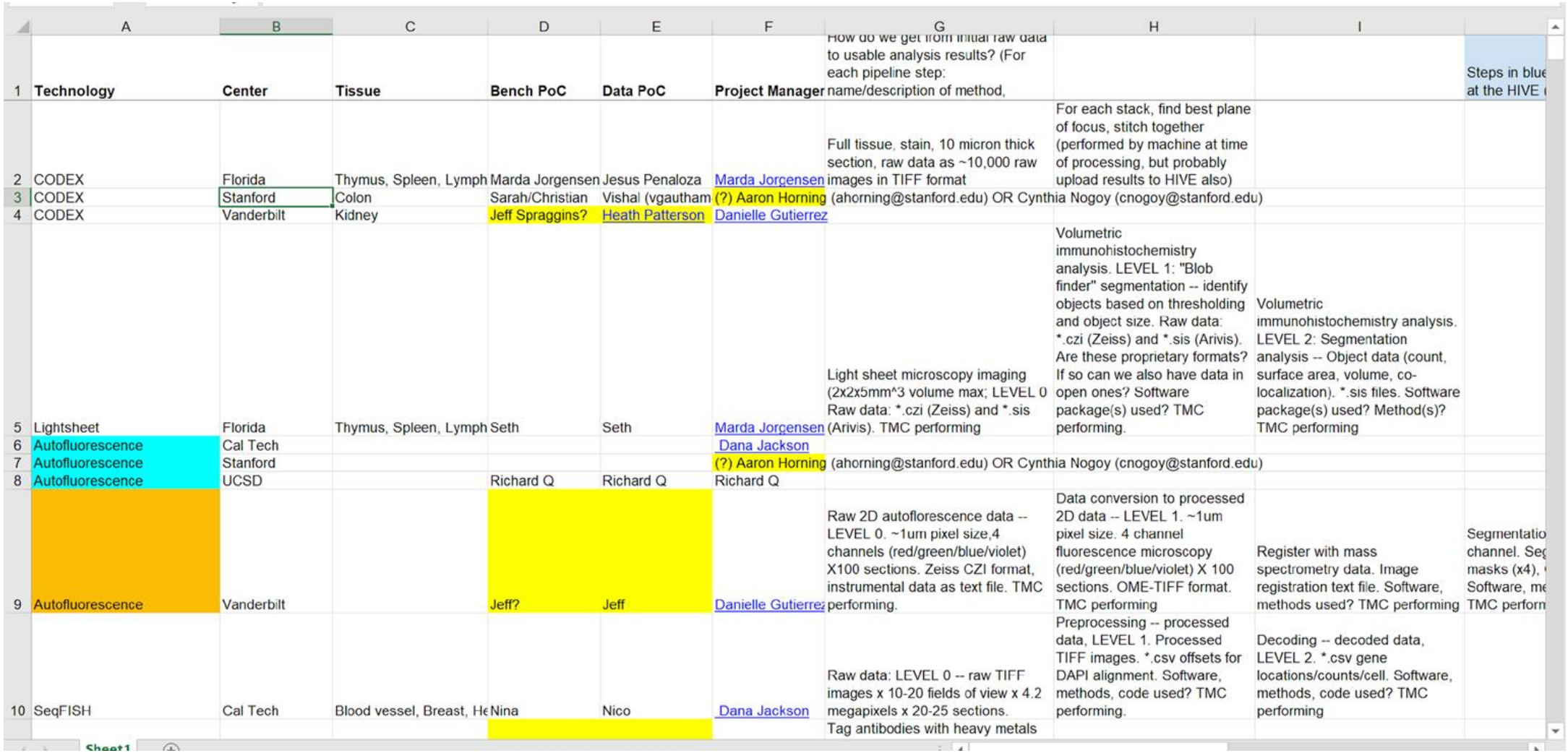
Center Nodes



Graph Legend

- Input File
- Output File
- Input Reference File
- Intermediate File

Microscopy Pipelines

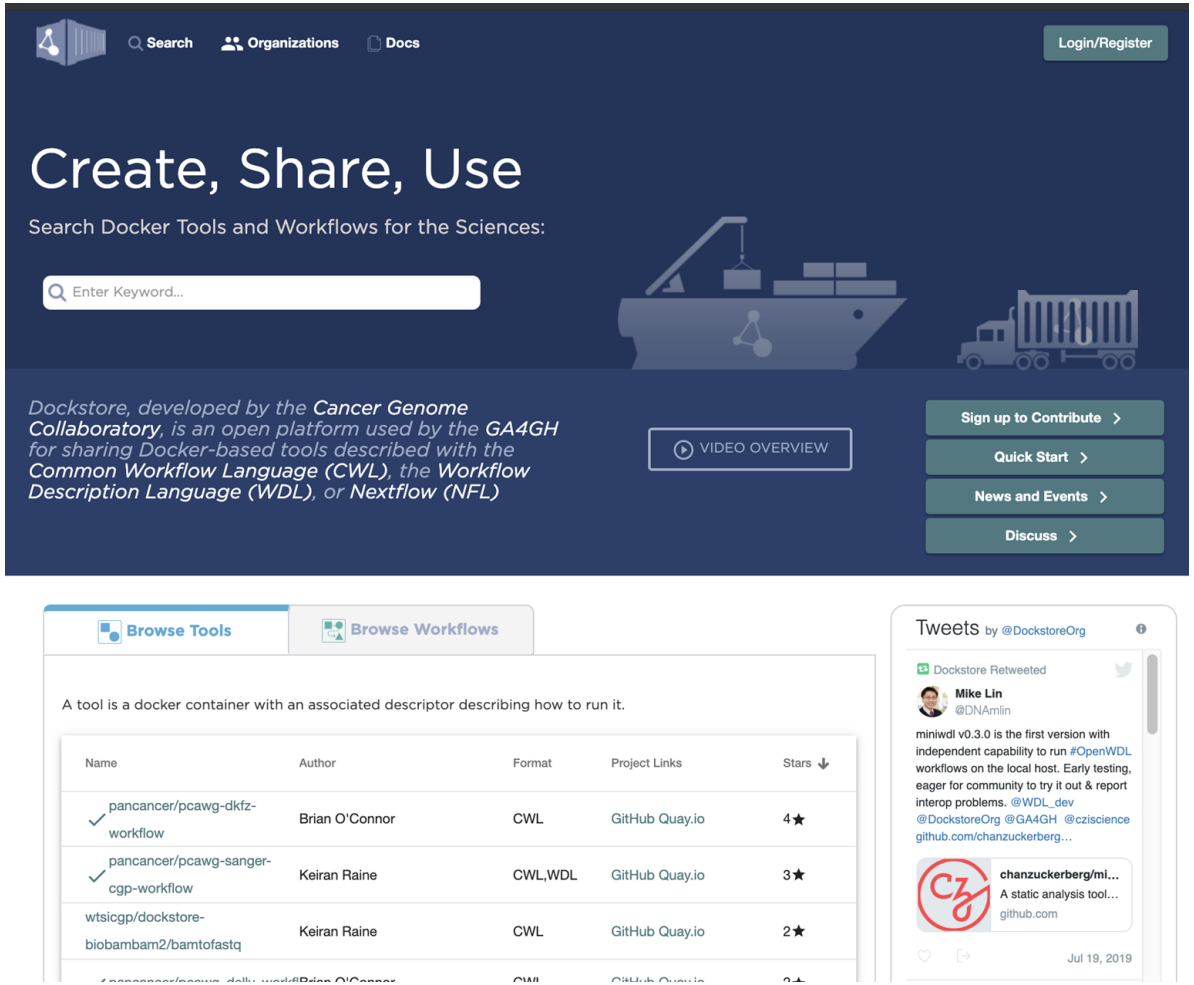


	A	B	C	D	E	F	G	H	I	
1	**Technology**	**Center**	**Tissue**	**Bench PoC**	**Data PoC**	**Project Manager**	How do we get from initial raw data to usable analysis results? (For each pipeline step: name/description of method,			Steps in blue at the HIVE (
2	CODEX	Florida	Thymus, Spleen, Lymph	Marda Jorgensen	Jesus Penaloza	Marda Jorgensen	Full tissue, stain, 10 micron thick section, raw data as ~10,000 raw images in TIFF format	For each stack, find best plane of focus, stitch together (performed by machine at time of processing, but probably upload results to HIVE also)		
3	CODEX	Stanford	Colon	Sarah/Christian	Vishal (vgautham	(?) Aaron Horning	(ahorning@stanford.edu) OR Cynthia Nogoy (cnogoy@stanford.edu)			
4	CODEX	Vanderbilt	Kidney	Jeff Spraggins?	Heath Patterson	Danielle Gutierrez				
5	Lightsheet	Florida	Thymus, Spleen, Lymph	Seth	Seth	Marda Jorgensen	Light sheet microscopy imaging (2x2x5mm^3 volume max; LEVEL 0 Raw data: *.czi (Zeiss) and *.sis (Arivis). TMC performing	Volumetric immunohistochemistry analysis. LEVEL 1: "Blob finder" segmentation -- identify objects based on thresholding and object size. Raw data: *.czi (Zeiss) and *.sis (Arivis). Are these proprietary formats? If so can we also have data in open ones? Software package(s) used? TMC performing.	Volumetric immunohistochemistry analysis. LEVEL 2: Segmentation analysis -- Object data (count, surface area, volume, co-localization). *.sis files. Software package(s) used? Method(s)? TMC performing	
6	Autofluorescence	Cal Tech				Dana Jackson				
7	Autofluorescence	Stanford				(?) Aaron Horning	(ahorning@stanford.edu) OR Cynthia Nogoy (cnogoy@stanford.edu)			
8	Autofluorescence	UCSD		Richard Q	Richard Q	Richard Q				
9	Autofluorescence	Vanderbilt		Jeff?	Jeff	Danielle Gutierrez	Raw 2D autoflorescence data -- LEVEL 0. ~1um pixel size,4 channels (red/green/blue/violet) X100 sections. Zeiss CZI format, instrumental data as text file. TMC performing.	Data conversion to processed 2D data -- LEVEL 1. ~1um pixel size. 4 channel fluorescence microscopy (red/green/blue/violet) X 100 sections. OME-TIFF format. TMC performing	Register with mass spectrometry data. Image registration text file. Software, methods used? TMC performing	Segmentatio channel. Seg masks (x4), Software, me TMC perforn
10	SeqFISH	Cal Tech	Blood vessel, Breast, He	Nina	Nico	Dana Jackson	Raw data: LEVEL 0 -- raw TIFF images x 10-20 fields of view x 4.2 megapixels x 20-25 sections. Tag antibodies with heavy metals	Preprocessing -- processed data, LEVEL 1. Processed TIFF images. *.csv offsets for DAPI alignment. Software, methods, code used? TMC performing.	Decoding -- decoded data, LEVEL 2. *.csv gene locations/counts/cell. Software, methods, code used? TMC performing	

What is Dockstore?

Dockstore is a free and open source platform for sharing scientific tools and workflows. It is a registry of Docker-based resources described using popular workflow languages CWL, WDL, and Nextflow.

- **Portability**
  - Run workflows in any environment that supports Docker
- **Interoperability**
  - Standardize computational analysis through GA4GH APIs
- **Reproducibility**
  - Create, Share, Use
  - Containers + Popular descriptor languages



Now on version 1.6.0, first presented version 1.25 at BOSC2017

OICR Ontario Institute for Cancer Research

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

The Human BioMolecular Atlas Program

The vision for the Human BioMolecular Atlas Program (HuBMAP) is to catalyze development of a framework for mapping of the human body at high resolution to transform our understanding of tissue organization and function.

HuBMAP@mail.nih.gov

https://commonfund.nih.gov/hubmap

Collections Members Events

HuBMAP Analysis Pipelines

Analysis pipelines developed by the Human BioMolecular Atlas Program

HCA Skylab Pipelines

Secondary analysis pipelines for the Human Cell Atlas.

nf-core Pipelines

Bioinformatics analysis pipelines used for RNA sequencing data

HuBMAP

The overall goals of the NIH Common Fund Human Biomolecular Atlas Program (HuBMAP) are to (1) accelerate development of the next generation of tools and techniques for constructing high-resolution spatial tissue maps that quantify multiple types of biomolecules; (2) generate foundational 3D tissue maps using validated high-content, high-throughput imaging and omics assays; (3) establish an open data platform that will develop novel approaches to integrating, visualizing, and modeling imaging and omics data to build multidimensional maps, and making data rapidly findable, accessible, interoperable, and reusable by the global research community; (4) coordinate and collaborate with other funding agencies, programs, and the biomedical research community to build the architecture and tools for mapping the human body

The Human BioMolecular Atlas Program / HuBMAP Analysis Pipelines

Analysis pipelines developed by the Human BioMolecular Atlas Program

github.com/mruffalo/cellranger-pipeline

Last updated Sep 17, 2019

Remove

github.com/mruffalo/chromvar-cwl/chromvar

Last updated Sep 3, 2019

Remove

Select edit to add a markdown description

Deliverable for Year 2

Goals for Year 2:

- We will work with the TMCs to make sure that pipelines needed to process HuBMAP data are available
- We will containerized pipelines for processing HuBMAP sequencing, imagining and mass spec data
- We will attempt to standardize pipelines currently used by the different groups
- We will work with others to make sure that the HuBMAP data is accessible, searchable and downloadable
- We will attempt to test joint analysis pipelines for HuBMAP data

• Current needs

- Cooperation from TMCs and DRTs

Collaborations

- We are starting to work on a project for the analysis of spatial transcriptomics and proteomics data
- Very early stages, no updates yet

What should HuBMAP Do ?

- Upload sample data for all expected modalities in the initial release ASAP
- Make sure that all the data you expect to provide is covered by one of the DRTs
- Better discussion and collaboration with external efforts (NIH or others)
- Start uploading actual data collected