Each group will have 12-minutes to present to the group. We will be collecting feedback for you from the consortia after your presentation. We recommend therefore that you focus your presentation time on areas in which you might like feedback and new thinking. There is a template for your report back here. Please place your report back presentation here.

This time is for each team to use as they need. You will have members from the Portal team joining your group. The below are suggestions only, please edit as you wish.

- Establish internal deadlines
- Work towards finishing any joint work e.g., are the data levels the same across multiple assays? Work on defining QA/QC for each assay and harmonizing across, etc. Work on harmonizing file formats (consider embedded metadata in images and other assay-specific questions)? Are there minimum information standards in your community that you want to adopt?
- Transferring data/metadata
- Getting UUIDs and how to use these in the metadata files.
- Transferring software
- How to execute your pipeline on the HIVE system
- How to know that the data is “ready for release”?
- What curation activities do you expect the HIVE will undertake?

Goal	% Done
Identify Assays & Centers & Reps	✓✓✓✓
Specific workflow for assay at each Center	
Definition of data levels	
File formats defined	
Assay metadata & file format defined	
Processing pipeline defined	
Identify potentially common processing steps & who/where it will run.	
Assay & data QA/QC criteria	
Understand how to upload data/metadata to HIVE	
Understand how to validate and transfer processing pipeline to HIVE.	

Action Items -

Notes:

Bill Murphy:

Having a common manner in which HuBMAP TMCs provide spatially resolved molecular distributions (including mass spectrometry imaging) will greatly facilitate analysis, comparison and search across images from different tissue sources, modalities, and centers.

All spatially-resolved molecular information resulting from a contiguous region of a single sample at a given spatial resolution shall be provided in a single OME TIFF file. Each file must have all required OME TIFF fields (e.g., pixel/voxel sizes).

A separate OME TIFF will be provided for each resolution. If the same sample is measured at different spatial resolutions, a text file will be provided containing the names of each of the OME TIFF files and the relative positions of the corners of each file in microns (or equivalent) to the first file.

- Each file will have one channel per molecular species or detection method used
  - Each species for IMS, or any multiplexed fluorescence method
  - Autofluorescence
  - bright field
  - phase contrast
  - H&E staining
  - Etc.
- All channels will be registered/aligned/warped to each other
- Channels will be named with HuBMAP standardized terms (e.g., protein/gene name)
- Additional channels may contain masks (as indexed images) to identify the locations of particular structures.
- Mask channels will be required for nuclei and cells. These will have standardized names. Other masks are optional. Additional standardized names may be defined (e.g., BloodVessel).
- If a non-standard name XXX is used, a tag XXX_definition containing a description) must be included in the XML portion.

An alternative to OME TIFF is to use a TIFF file with a HubMAP XML header (similar to OME XML) but with the important difference that it can contain channels with different spatial resolutions. In that case, the HuBMAP XML will contain tags to define the corner positions of each channel relative to each other.

Microscopy DRT metadata session from Sept 2019 HuBMAP meeting

Created 9/25/2019

Clive W, Peter K, Elizabeth

Metadata options collected

The following text is a list of every post-it note collected during the Microscopy DRT metadata session, specifically seeking to define metadata for the data release effort.

Ab Clone

Titer

Priming Conjugate

Fields of view

Pixel size

Thickness of slice

Part of tissue

DAPI

Cell Count

Regions of interest & spatial annotations (nuclei, cytoplasm)

Lamp intensity, power

Microscope settings

Exposure time

Gain

FISH probe set / probe library

Probes

Readouts

Fluorophores

Channels

Antibody clones

Channel order (marker detection)

Antibody used (RRID, clone, lot, expiration, validation)

Antibody metadata

Antibody validation

IMC antibody

Ab: conjugate chemistry label

Ab: clone, titer, vendor, lot

Intra run staining controls

Autofluorescence intensity values

Markers

Channel (Spectra)

channels / channels per round

Cyclic metadata - staining order

Exposure time

Exposure time (cycle #)

Magnification

Objective / pixel um conversation (detector -> microscope)

# of pixels

Ab clone

Ab titer

Vendor

Vendor lot

Conjugation method

intra run tissue controls

For cyclic methods - cycle order

Cell segmentation & single cell table

Segmentation markers

Guidelines for low level data processing

Data: raw

Data: processed

Post imaging processing

Processed & segmented data

FCS files for analysis

Experiment metadata

Preservation: FFPE vs Frozen

Assay data volume (how much data comes with the assay / metadata)

Interpretation

IMC: Peak call transform criteria

Parameters of hybridization

Quantitative data format

Epitope retrieval conditions General protocol

CODEX - all using same software? Processing?

IMC: mass spec para

Hybridization rounds

Number of Z-stacks

Exposure times

Create location send to hive

Assays:

- Slide Seq
- Cell Dive
- DartFISH
- CODEX
- Light sheet

Feedback, questions, and response to report-back session

? - Imaging standardization of metadata might be difficult - recommend setting & agree upon small set of common metadata ____a platforms

O - need to align decision to cou_____ to OME-TIFF w/ Jonthan / IEC? Not what he is planning?

? - could a single point of contact from HIVE be identified?

? - Articulate the scientific reason for HIVE extracting info from OME-TIFF

? - Who from the HIVE will be the point of contact for this group?

? - When you say you need more reps from HIVE, in what specific expertise are you looking for?

? - How can you more strongly get TMCs to give you data?

O - Have you checked with other consortium we ____ ___ imaging metadata?

? - plan for reproducibility?

? How are you harmonizing specimen collection, stabilization, and preservation?

? - how will you decide which data analysis pipeline to use for CODEX? What about mkgration w/ scRNA or ATAC?

O - consider working toward FAIR preprocessing using Docker containers with workflow specification in CWL, WDL, or NextFlow and upload to Dockstore

? - What preprocessing of data is done prior to sending data to HIVE as OME-TIFF? How to ensure those preprocessing steps are FAIR?

+ - Good timelines

+ - Excellent organization of wide breadth of data

+ - Articulate, great sense of humor

+ - finding minimum requirements to advance

+ - exchange information with HIVE to verify that HIVE is extracting the right information

+ - getting more people involved

+ - Imaging “Fresh” / new person presenting - good presenter [Congrats, Elizabeth … :)]

+ - thank you for acknowledging that this is hard

+ - liked aggressive 9/27/19 timeline

+ - back and forth metadata extraction & testing

[] - HIVE has a minimum metadata & data formats document

[] - HIVE will determine if TIFF file format is “correct”

+ - Testing is great as early as possible

+ - great use of deadline! Fri Sept 27!

+ - HIVE evaluation of data sets

? - Have we settled on ontologies to be used in annotations?

? - are there controls for CODEX et al?

? - What level of annotation is the goal?

? - Have you considered common verses individual data processing, i.e. will this happen at component sites or is it a HIVE responsibility?

? - What are more specific items “in process”?

? - could a set of questions be defined for each modality that drives metadata lists?

? - for timeline, how may you expect HIVE to evaluate datasets?

[] - The meeting w/ HCA will be too late to effect the first data release

? - Imaging validation positive and negative controls

Microscopy Breakout – 9/24/2019

Attendees - Kevin, Elizabeth, Kayvi, Tubin, Nils, Shin, Nina, Nico, Qian, Chen, Mike A, Shannon, Dena, Maria, Tyler, Marda, Sinem, Yousef, Harry, Stephen, Zorina, Kun, Anup, Peter, Clive, Jonathan, Richard, Lucy, Bill M, Jim S., Jason Swedlow, Pehr, Ziv, Neil, Stephen H

Clive – where are we going with metadata, QC, where is the handle for the HIVE goes
CODEX, light-sheet, seqFISH, Cell Dive, DART-FISH are the modalities we have in the DRT
I don’t know if we’re going to have all of these modalities on every tissue by the data release. We’ll probably do spleen, but we have to prioritize, which of these modalities need to be ready by December
Yousef – MD should be compatible
Zorina – want to start planning now how we are going to get to 3D
Mike – what’s the definition of raw versus processed?

Clive – we’re going to decide that today

Mike – general guidelines for post-processing?

Clive – yes. The HIVE should have a standardized way of what we're sending them. But we might send our own processed data along with the raw.

This is where we are, it's a long way to data release. The feasibility of this depends on the surgeons sharing their tissues with us when they get it

How we're getting the assay data from the HIVE, this is what we have to figure out

Mike – we have data sets that are fully annotated from MIBI

Harry – recombinant Abs, and top-down mass spec, which is different from the MS next door

Richard – Slide-seq

Jonathan – any of these where the data isn't images? Slide-seq?

Mike – do you know when the data would go live?

Jonathan – June 2020, but your data needs to be in by December

Mike – okay, but doesn't go anywhere live until June? Yes

Clive – need to prioritize the methods

Jonathan – any of these that won't be done?

Anup – Cell Dive won't be ready by December

DART-FISH, no

How ready does something have to be? And how much do you need?

Sufficient to contribute to a data release

Not just one off

CODEX and seqFISH were promised – they'll be there

MIBI – have data sets that are ready to go – have raw and processed

Clive – is the HIVE ready to deal with raw MIBI data?

Maria – just learning about it now, so I need some time

Mike – we have the raw data, then there's a mask that shows where the cell is. We have patient data.

Jonathan – you have enough provenance data so we can meet some criteria?

Mike – these are all clinical triple negative – 42 patients

Harry – do you know where the tissue came from? The clinical data is different

Zorina – are we going to have metadata and a picture of the tissue? That will be essential

> The pathology report? Maybe, but the pathology report could say it's normal, and the assays say no. An H&E slide-scan? Yes.

Anup – normal tissue?

Mike – have normal controls, have a cohort of triple negative breast?

Anup – are you interested in diseased tissue?

Richard – we're interested in normal, but if we have data from diseased tissue, nothing says we can't house it. Generating data though should come from normal data. It's a good example of where we want to go.

Harry – should be stored on HuBMAP? Or can be part of the released?

Richard – the money we gave you shouldn't go to analyzing diseased tissues, but if you have data, we'll be happy to house it

Ziv – but we don't have any metadata plans for diseased tissues, nothing is in place for that

Mike – if it isn't in the scope of what you want, that's fine, we can send later

Jonathan – so, we're going to exclude things that are part of DART-FISH, Lightsheet, Cell Dive, Slide-seq? We have to make sure we're still on track

Important to keep these needs in mind, but we need to focus on the needs for the methods we are going to focus on?

Yousef – I think the metadata is applicable to make different techniques

Mike – the metadata will be fairly generic

Peter – what's the data and metadata needed to be collected that's specific for those methods?

Nils – the difference between the metadata between the methods – there will be differences between them

Sinem – good to include Ab information in the metadata

Kevin – do you guys do EC50 titering?

Mike – yes.

Clive – all of them? For every run?

Mike – yes

Antibodies – positive/negative controls

What's a negative control?

Mike A – dependent on the protein target. Can have a Consortium definition, or leave it to the sites to decide. Best controls have known positive/negative in the same section. The Consortium should have a high bar of what the control should be. Have to get that it's staining the right target.

Harry – peptide-based Abs that you block – is that a good control?

Richard – can say this is a beta release, and that we will continue improving as time goes on

Mike – you need interrun staining controls. Can use some of the spots in the normal run.

Nina – so we're either going to look for marker genes for each tissue, or probe for same genes to get colocation to localize the signal.

Clive – RRIDs, way to capture what we're doing

Jonathan – are people using them?

Clive – not a validation, just a way to label the data

Jonathan – right, but we have to identify things

Mike – the ab clones are central, and are known what they bind too. For the top 150-200 clones, they've been studied a ton.

Anup – having the information about the antibody clones are very important

Sinem – traditional FISH they use scrambled nucleotide sequence that won't bind to anything

Zorina – there should be a gold standard for controls

Mike – yes! But there isn't

Clive – okay, let's move on

Have to assign the channels

You are barcoding on seqFISH – you collect one channel at a time? Or you have an order

Nico – now we do it by channel to not have to account for differences

Clive – and it's all ingestable by the HIVE

Jonathan – sure, as long as everyone is using a similar JSON, or you have a tool that can extract it into JSON for us

Elizabeth – microscope settings, lamp intensity, number of cycles, detector, detector gain, type of camera, clearing agent,

Harry – those might be a bit too much detail, do you want to collect it?

Jonathan – is this good data? Do I have a novel way to analyze this, and the information I need to analyze, do I have the information I need to reproduce someone else's work? We can take all of this

Harry – but we don't collect all this now, do we need it? None of this

Jason – all the md comes in a tif file that you have extract

Jonathan – may as well just give us the data in the file, or the tool to extract it

Number of Pixels in an image – field of view

Probably fundamental rules about what you need to make something an image

Jonathan – can you guys give us the specs and we pull out the data? Is it a hard step to say what something is, and you guys just need to send us those files? We can do that and get MD moving before we start moving. And if you send it, we now have something we can QC. If you tell me "this file should have all of this information" and it doesn't, it will be a signal to us

Elizabeth – I've made an R script to pull out the important metadata from the 20K lines

Harry – if you know what information you want, that makes it easier for us

Richard – experience has taught us anything that you can export off the machine is the best way to go. The less manual entry is better

*******Get tif files from each of these machines and see what we need *******

Mike – do the vendors tell you the header structure?

Jason – they write the specification, but you have to get in to see what's in there

Jonathan – they should convert before they send it to us? Find what people might use. Need to say what we're looking for.

Ziv – need to know what data is coming

Clive – we have a commercial version of CODEX, you have a home grown version – send the file to Peter

*****make a globus location for people to drop their files in to see what they can extract out

Jim – there's a validation tip to run to see if anything comes out

Breakout #2

Clive – have a bunch of questions from before – we'll look at the questions and comments from before

Jim – we don't have to spend a long time discussing metadata – what you got? Send it

Setting the next DRT meeting – Friday Oct 11, 2019 at 2:00pm EDT

Elizabeth, Peter, and Clive will go through the feedback

We'll leave MD alone – do we want to talk about QC

Maybe define levels of QC? QC of the data, and the assays, and the HIVE

The group that generates the data has to do the sanity checks, because the HIVE can't say you did the experiment wrong

Ziv – it depends on the data – what if you upload RNA-seq data and nothing aligns that means there's something wrong there, or you upload images that are black, there's something wrong

Clive – where is segmentation getting done? At the TMC? Or the tool provider?

Jim – haven't talked about data analysis at all

Clive – what point are we going to get to by December 31

Marda – for CODEX - if everyone submits into one pipeline, that would be the best way. If you submit in aquioa all that data is there – you get a JSON, a TIF, a text and list all the data they have –

Ziv – we would need for you to tell us what data you need. If this will be the only processing done, we have to decide. But the starting point for us is which pipeline you decide to use.

Clive – Garry Nolan is going to provide his format

Marda – the raw data will look the same. If you use Aquioa you get all the files you need – they can do all the processing you need to clean it up. If you want to run your own processing you can – this way you'd have a seamless pipeline for CODEX.

Someone said they would have CZI files – that will be an entirely different system – and that could be a minor software

Richard – we need analyzed data alongside the raw data.

Ziv – do we want to end up with genes/proteins/cells? But it's not going to be comparing them.

Stephen – you are working in a space that's actively moving. At some point you'll have a process data stack, the originator's first pass. After that you can figure out how to work downstream. Can we go back and work together organ by organ, and then you can.

So, you will be uploading two files – raw data and processed data

Sinem – after the registering and stitching, I think that's the better form of raw data

Stephen – I think should stick with raw for now – maybe in the future

Phil – have you seen the processing spreadsheet?

Clive – yes, we have that, we're on the boundary of raw and processed. Seq-FISH does their own

Stephen – you're going to stratify raw data to secondary raw data – there are some processes you're going to want to go throw

Harry – unstitched data?

Stephen – yes

Ziv – if you give unstitched data, that's up to you

Harry – we have a section of tissue on the slide – putting the tiles together should be the raw data

Elizabeth – we're just start with the raw and work from there

Harry – what's common between Aquioa and Nolan?

Vishal – Nolan has some customizable features, has tile registration and stitch dimension. You can process 4-5 at a time. Aquoia will only let you run what at a time

Use Aquioa that's "process only" without any JSON or files to be used for analysis.

If you have 22 cycles, with 6 regions, and 7.9 tiles, the data's going to be huge

Ziv – this is commercial software which we haven't discussed – they are Windows based, so we can't run them at CMU, but that's another problem

Stephen – I'd love to have the stitched data, just put it in a TMA

Yousef – sounds like you are just designing the whole system around CODEX, all of those JSON files is specific CODEX – what do you propose for people who don't use CODEX?

Stephen – send us raw data

Harry – so we should just send raw data?

Ziv – no, we want processed data too, not everyone raw data

Vishal – I think we should have processed data, run on your analysis tools

Ziv – so both, but if we can't do both, which one? Are we doing the processing ourselves? Or is that TMC? All the CODEX data is processed with the same pipeline, which is better than we have four separate pipelines

Yousef – so we'll need patches for each?

Sinem – if that's the data then maybe a common algorithm? Because then we have to worry about the naming convention.

Stephen – that becomes a higher md system. The tiles should be labeled by the system.

Ziv – seqFISH – you have a pipeline for spot coding. Is there a stage where what you're doing converges with what the CODEX people are doing?

Kevin - You need a cell x gene. We do segmentation manually to ensure high quality

Nico – we use a pixel classifier, it's semi-automated training

Ziv – we are looking for places that they converge. Is there any place in the modalities we can converge?

Richard – the cell x gene table, a polygon that represents a cell.

Ziv – the segmentation could be done by the same method

Kevin - But we aren't happy with it

Ziv – sure, but we need it to scale

Anup – you are nuclei staining with DAPI

Phil – we can host all of the different version of the data

Clive – so three data sets, raw, each TMC processed, and HIVE processed

Richard – then you need to give the HIVE the code you are using

Ziv – yes, there is software that's available

Richard – there are many pipelines for doing segmentation that's open source. There are many different ways and workflows for doing segmentation.

Gloria – ontologies, what are we using for them?

Ziv – that's a md issue that will come later

Jonathan – ontologies are reference terminologies. To force individual labs to use a particular ontology forces them to do a lot of work. We can translate it without making people do all that.

Richard – we don't want to annotate all our cell types in all our data sets

The reference for what's a cell type and what's expressed is a small matrix

Jonathan – we don't know how many cell types there are.

Jim – there are so many of them there's no point in naming them

Ziv – we're talking to HCA continuously. They have decided to stay away from cell naming

Harry – CODEX what do we want to provide?

Clive – raw and processed. Raw tiles, and processed and you get a TIF and a JSON file

Jonathan – will we require an overall view of the thing?

Harry – we have a system where you can do H/E, IMC, autofluorescence, whatever

Jonathan – we want an overview image of the thing – a low magnification. Is that normally done?

Marda – no.

Jonathan – do we believe that we need an overall image, and if we do, do we believe it can be done in all of the labs

Harry – if you are doing 16/18 abs, and 5 washes, you might not end up with what you think you have

Elizabeth – we've done autofluorescence after CODEX and it's worked out fine.

Clive – good quality data, ready for release and then the HIVE can process

Jonathan – are we going to find that the information we want is in the least common denominator of the headers? Or will we find that they don't have a lot/anything in common?

I'm happy to support this process

Clive – we have the ISHes to deal with

Nina – we have the separate data – raw and processed images

Nico – our images are in Z stacks, is that okay

Richard – that's what you feed into the pipeline, so is STARFISH. We want the processed data.

Nico – you want an overall picture?

Sinem – but even that will need to be stitched

Richard – if we had raw and processed in the HIVE, we can try to process

Seth – lightsheet – really big Z stack. I just need the smartest markers you have.

Harry – he's got 3000 images to provide.