Imaging Mass Spectrometry : HuBMAP

Knowledge Transition document

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

Preface & Credits:

The document is written based on self-learning, question & answers from forums (so please kindly excuse typos/errors and feel free to correct as required). Mass Spec techniques, analysis is for the most part self-learned, can be updated with better information as-you-go. Cytokit related details in this document are updated based on discussion, guidance from Maria Keays, who is the point of contact for Microscopy at HuBMAP from Sanger Institute. Mass Spec analysis using PCA and NMF were updated as discussed and guided by Dr. Matt Ruffalo, who is the point of contact for Sequencing and overseeing Mass Spec pipeline work. The Mass Spec Imaging analysis is headed and guided by Dr. Robert Murphy. Mass Spec pipeline work is under immediate supervision, guidance from Prof. Ziv Bar Joseph. Thanks to everyone for guidance. Grateful for the support and guidance.

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

Introduction to Mass Spectrometry:

Traditional Mass Spectrometry is used to measure the mass-to-charge ratio of ions that can be isolated based on mass of the elements.. This technique was initially used to isolate isotopic mass of the elements. Refer to Mass Spectrometer [1]. The same technique, with variations, is used for isolating and recognizing the m/z ratios with intensities that results in a mass spectrum. There are different techniques such as MALDI (Matrix-assisted laser desorption ionization), LC-MS (Liquid Chromatography - mass Spectrometry).

What is MALDI ?

MALDI(Matrix-assisted laser desorption ionization) is most commonly used in the study of metabolomics and proteomics [2] . For a quick video on technique, refer [3]. There is a special process to contain the tissue sample and depending on the type of analysis that has to be made, a number of steps are required to prepare the sample. Quality of analysis using MALDI depends on sample preparation, refer [4] and since HuBMAP's Vanderbilt team uses a specific workflow that uses Time Of Flight technique (MALDI TOF), sample preparation in the [4] refers to MALDI TOF technique. The intensities are total ion counts of a given mass-charge-ratio based on time of flight readings in the experiment.

What is LC-MS ?

LC-MS (Liquid Chromatography Mass Spectrometry) is often used for metabolites. Vanderbilt uses workflow that performs both MALDI + LC-MS.

Vanderbilt's workflow and details:

Vanderbilt's workflow [5] is precise, however, for data that will be released by Vanderbilt, is an image file (i.e. intensities for X,Y so each slice is a metabolite and its corresponding intensity i.e extent of its relative presence in a given x,y coordinate). We only focus on images. Some of the preliminary analysis was done from September 2019 Vanderbilt's data. Details in the following section.

For details of specific processes followed by Vanderbilt refer [6].

Quick details of each of the steps from Vanderbilt's workflow:

1. Landmarks, topographical sectioning to preserve dorsal, ventral and temporal poles. Eg: mouse retina -> there is a separate sectioning protocol depending on the type of tissue provided and need to be sectioned based on a protocol to preserve the biological properties and landmarks of biosample.

2. Liquid Chromatography Mass Spectrometry is performed on sample types that are thermally unstable, large, polar, ionic or non-volatile. Examples: nucleotides, peptides, steroids, hormones, dyes, fatty acids, acids and alcohols.
3. Defining IMS measurement regions and sample preparations are required to ensure the integrity of the sample that is being studied. Example: if a tissue sample has mass spectral data that can be registered to spatial locations from Microscopy, then having a matrix of such spatial location information ensures spatial integrity of sample. So it is because of this reason, mass spectrometry cannot completely define, provide regions of interest (ROI) for metabolites/proteins from mass spec images alone.
4. Intensity of an ion in a coordinate system represents the relative position of mass spectral acquisition from biological samples.
5. imZML file format: This is an instrument generated data. This includes full mass spectral data. This data should have had intensity normalization, m/z alignment and calibration.
6. At present, the DRT (Data Release teams) at HuBMAP for Mass Spec, would release mass Spec data for 2D images only. However, we would eventually expect 3D images, say after June 2020.
7. Currently the current Mass Spec pipeline is for 2D Mass Spec data analysis only. (Referring to Github links).
8. Input to MALDI experiment, requires spraying and sublimation of the input sample sectioned from tissue sample. A spectra is collected at each pixel. An ion image is collected for each mass-charge-ratio i.e. m/z ratio.
9. Every pixel's spectra would have multiple peaks, but only significant peaks (that result from peak picking, peak alignment and peak filtering etc) would be provided in the mass spec data provided by Vanderbilt.
10. Tissue sections used for Mass spec and Microscopy - we assume are alternate sections are fed to mass spec and microscopy respectively. This eventually needs to be stitched back from segmented mask to mass spec image as part of the pipeline, which is being worked by Maria Keays at the moment.
11. Since tissue samples/sections used in Mass spec and Microscopy undergo some preliminary processing, they cannot be re-used and hence alternate sections go to each of respective instruments.

September 2019 Sample data from Vanderbilt:

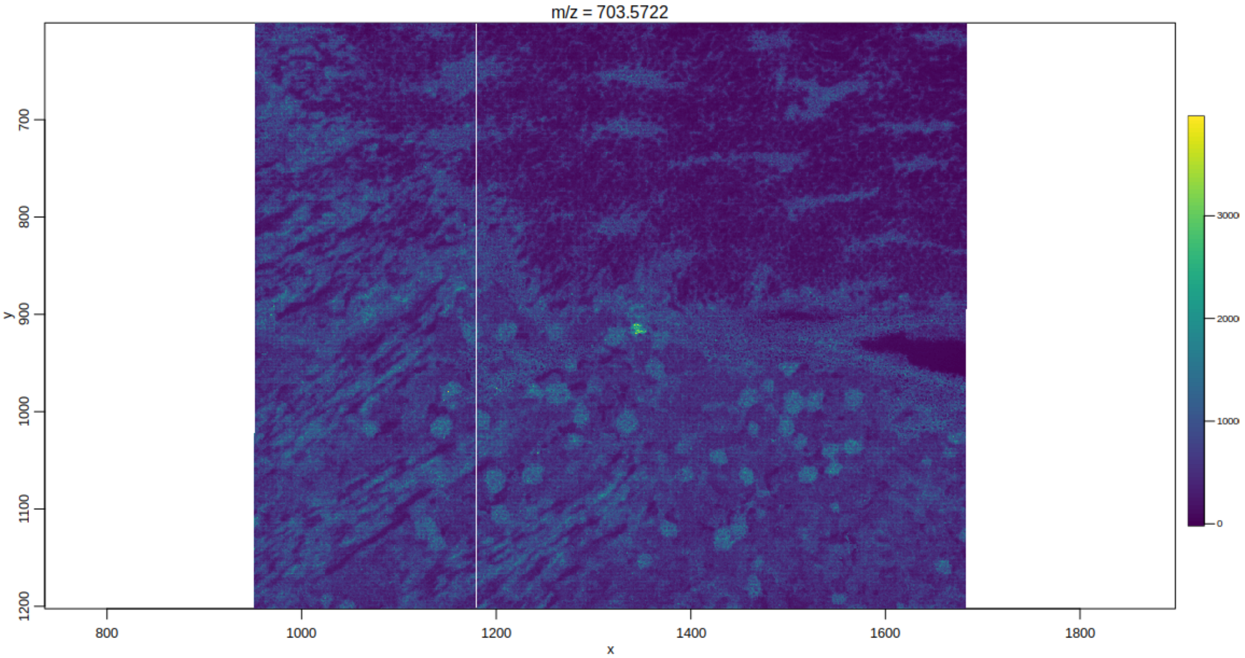
Summary of sample Mass Spec data:

The folder contains peak metadata and IMS data. The data is broadly categorized into : columnar, imzML and tif folders. Columnar data contains csv file with intensities for 47 metabolites i.e. m/z ratios. Each m/z ratio corresponds to a metabolite that is mapped in peak metadata file (which contains theoretical and empirical readings of m/z ratios). Often metabolites listed here in peak metadata would have assignment values with additional molecules, but would always contain additional molecules such as SM(d34:1)+H i.e. if you look up in https://www.metabolomicsworkbench.org/data/mb_mass_form.php would probably list SM(d34:1) and NOT SM(d34:1)+H (as here +H is possibly attached to original metabolite or this could also be a new metabolite not yet existing in present metabolite). This is just for reference, it also appears not all metabolites are already available in a single global database for public view. Example:

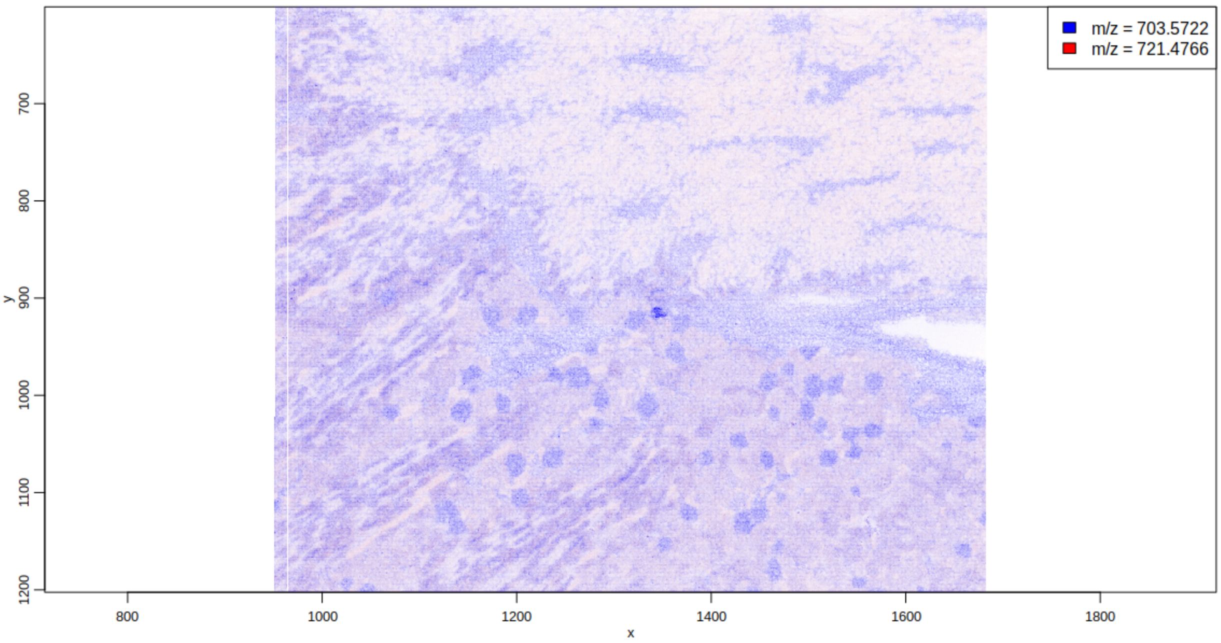
m/z	**Assignment**	**Elemental Composition**	**Theoretical m/z**
675.5366	DG(42:8)+H-H2O	C45H72O5	675.534687

So when mapping metabolite's composition to a corresponding name from, say metabolites database, it is easier to look for a combination of Assignment (DG(42:8)+H-H2O) + metabolite composition (C45H72O5) + theoretical m/z ratio (675.534687) together.

Plot single slice i.e. molecule in the entire grid: SM(d34:1)+H, C39H79N2O6P, 703.5722



Plot two molecules for X*Y : 703.5722, 721.4766 i.e. C39H79N2O6P, C39H71O8PNa :



Summary of analysis of Mass Spec data:

A preliminary analysis of Mass spectrometry analysis is available here: https://github.com/sushma-ananda/mass-spec . Mass Spectrometry data from sample, suggests intensities are normalized, with techniques such as peak picking, normalization, smoothening etc. Specifically the data that Vanderbilt provided is already pre-processed and normalized hence we do not have 100,000 number of metabolites with 400K pixels per metabolite and instead baseline reduced values with 47 metabolites and 733*602 # of intensities in final IMS data. We only receive normalized, peak picked, aligned values. For a given metabolite, from this dataset, intensity of ~600 to ~77000 TIC (The intensities are TIC i.e.total ion counts) is possible signifying its extent of presence in a given pixel i.e. x,y coordinate. Following analysis was performed on IMS data. Resolution of the IMS image is 10 μm. Some intriguing insights, summaries put together as follows:

- A principal component analysis, non-negative matrix factorization analysis was done. Since we would not expect negative intensities, having Principal components with negative values is not appropriate results for Mass Spec data for metabolites, NMF is a more suitable technique to study. This analysis is available in a notebook : https://github.com/sushma-ananda/mass-spec/tree/master/ims_analysis
- We can find correlation between metabolites or proteins from this dataset. For example, for the given pixel in tissue sample, how likely are Metabolites M1 and M2 can coexist ? Would M1 suggest the presence of M2 and vice versa or not are interesting questions. Colocalization is one method to find this analysis. This analysis, which is based on a rough pipeline worked out by Maria Keays from Sanger and was extensively modified to suit the current data from September 2019 is available at: https://github.com/sushma-ananda/mass-spec/tree/master/ims

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

- There are some pixels generally missing, which is an expected part of instrument generated data. For example, the sample data from September has significant number of intensities values that do not exist near the top right corner and lower left corner of 733*603 pixels mass spec image. We notice that 729 values are missing in this image. This often occurs since samples placed in a MALDI experiment might lose out-of-frame data something like if you scan a document and some parts of corner edges might get cropped /out-of-focus. So, for a better understanding, sample preparation technique recommended earlier in this document will provide a practical reference. Each matrix is filled with a sample that has been pre-processed. The ignorable precision loss that occurs to prepare Matrix-assisted samples for MALDI instrument laser desorption process seems to be common and is easily recoverable by Software tools, so far such loss is assumed to be ignorable. In case of September 2019 data, since 729 intensities are not available/missing, which is ignorable for a 733 * 602 image or can be assumed to be 0, seems a common pre-processing step for Mass data analysis in the online forums.
- IMS data received is already normalized: as per workflow - number of metabolites are 100K however the data we received has only 47 metabolites. The R tools such as Cardinal, EBImage support peak picking, peak normalization, filters, aligning and baseline reduction processes.
- For upsampling IMS images/slices to corresponding segmentation masks: upsampling from 10 μm resolution to 20x transformation requires finding/estimate corresponding intensities of each metabolite of 20 new pixels. For example, pixel 1 (x-970, y-603 coordinates) for metabolite (m/z) 1 has an intensity of 40,000 (total ion count). During upsampling, we would have 20 pixels which need to normalize based on considering mean on intensities of surrounding pixels, thresholding these values to roughly estimate intensities in upsampled data. Techniques for upsampling - yet to be identified.
- Lastly upon manual inspection, it has been found that a downsampled Microscopy image, can exactly fit in mass spec OME tiff image at the pixel location 952 and 601 on a VAN0001-RK-1-21-PAS-pas_toIMS.ome.tiff. For example, slice the Mass Spec image for first metabolite and you can exactly overlap it at the pixel location 952 and 601 on a VAN0001-RK-1-21-PAS-pas_toIMS.ome.tiff.
- Another approach to colocalization is: as we receive each tissue sample, lets say, kidney sample and we find colocalization of 47 metabolites (a 47*47 colocalization matrix). Then we receive new tissue sample, lets say, lung and we find 54 metabolites (another 54*54 colocalization matrix). We find that there are 81 new unique metabolites in both tissues. This time, we re-compute colocalization and find new 81*81 colocalization matrix for, let's say kidney. Thus this intersection of metabolites to build a colocalization of various tissues, would help get an analysis: for example, presence of metabolite M1 would infer presence of metabolite M2 in kidney tissue but presence of metabolite M2 may not indicate any information about M1. But in lung tissue, M1 would indicate presence of M2 and vice versa. So we would then have 81*81 for each tissue i.e. 81*81 matrix for kidney and another for lung indicating such correlation. The same, I believe, with minor modification, can be thought of about mass spec data for proteins, which is present in December 2019 data.
- Talking about colocalization: one example I came across was this method, where it was an attempt at building a gold standard data for colocalization of metabolites. This was a very interesting work: https://github.com/metaspace2020/coloc and specifically, "no learning" option is more relevant, given current (September 2019 data):

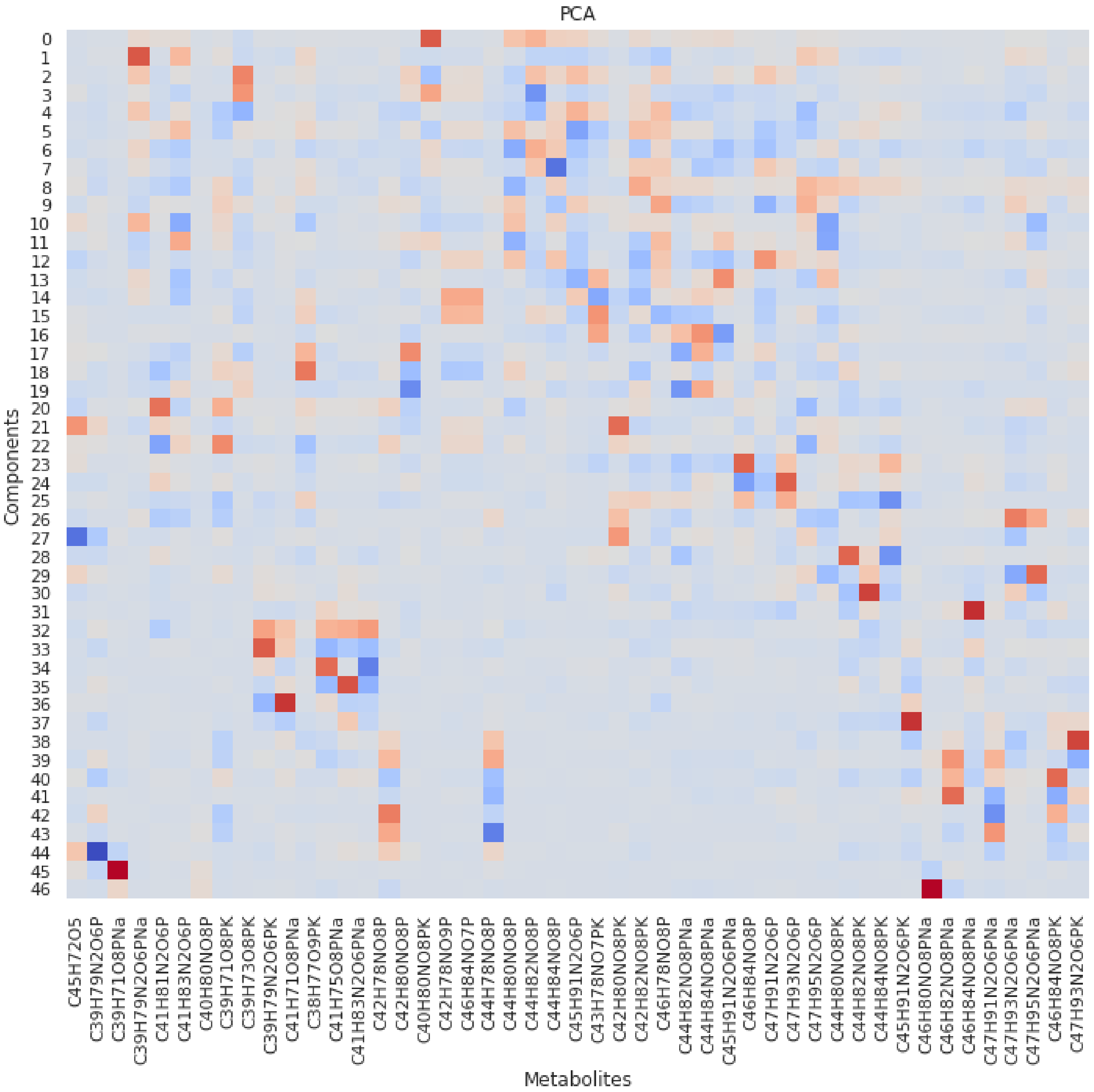
https://github.com/metaspace2020/coloc/tree/master/measures/NoLearning The same can be replicated for September mass spec dataset, however, would need to be run on the GPU for HuBMAP's Mass Spec data.

General Consensus:

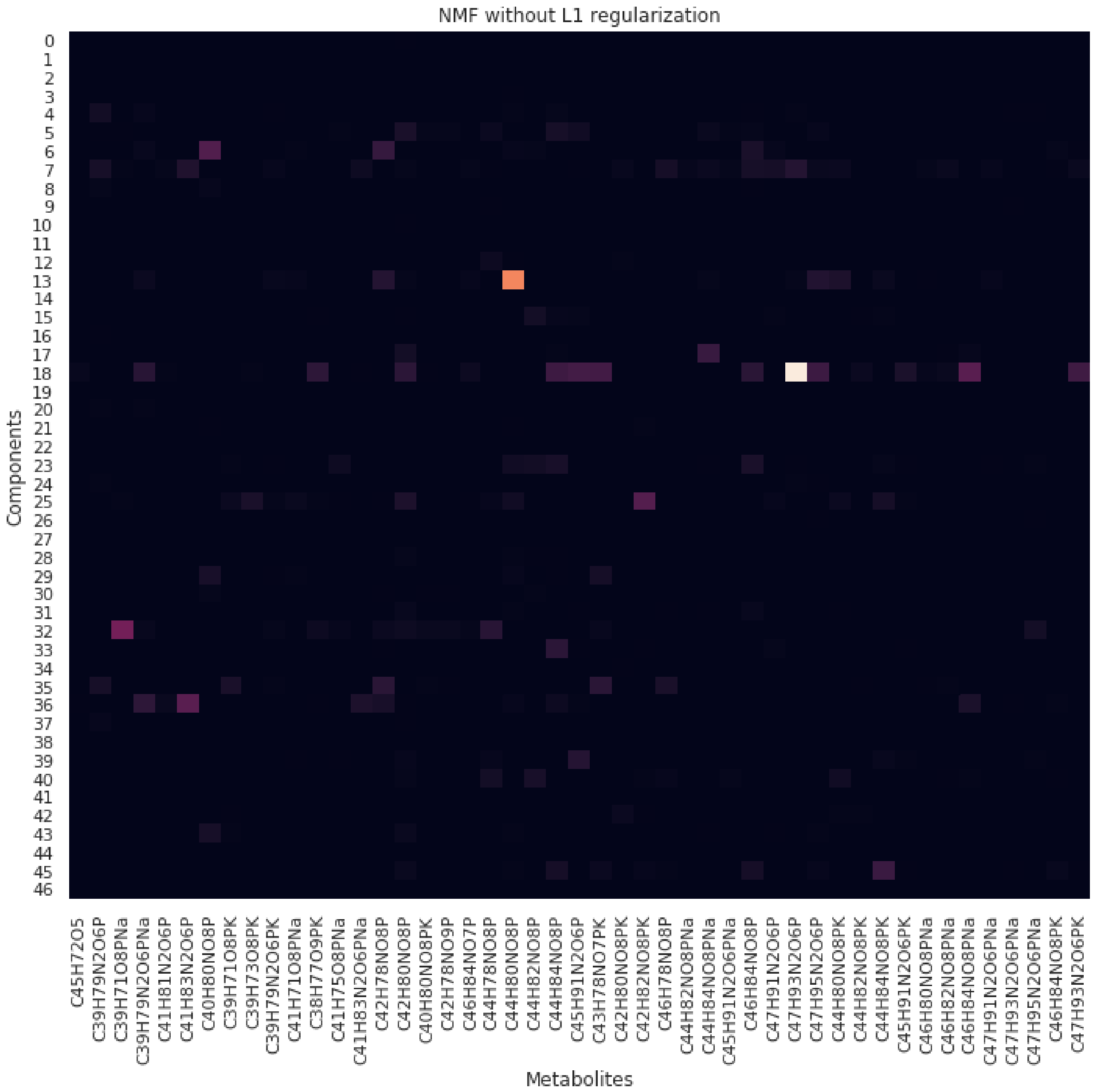
Imaging Mass Spectrometry data by itself, is not complete, given the goal of finding presence/extent of presence of metabolites/proteins in a given tissue requires the tissue sample to be mapped to corresponding segment/region of interest with reference to a Microscopy image of entire tissue sample. So depending on the segmentation masks obtained from Microscopy image's segmentation task to find cells, different parts of a tissue etc, IMS data needs to be upsampled (for this sample data: spatial transformation of 20 times is required) and then stitch the image to map IMS slices to get a complete metabolite distribution in a cell/various parts/ROI of tissue.

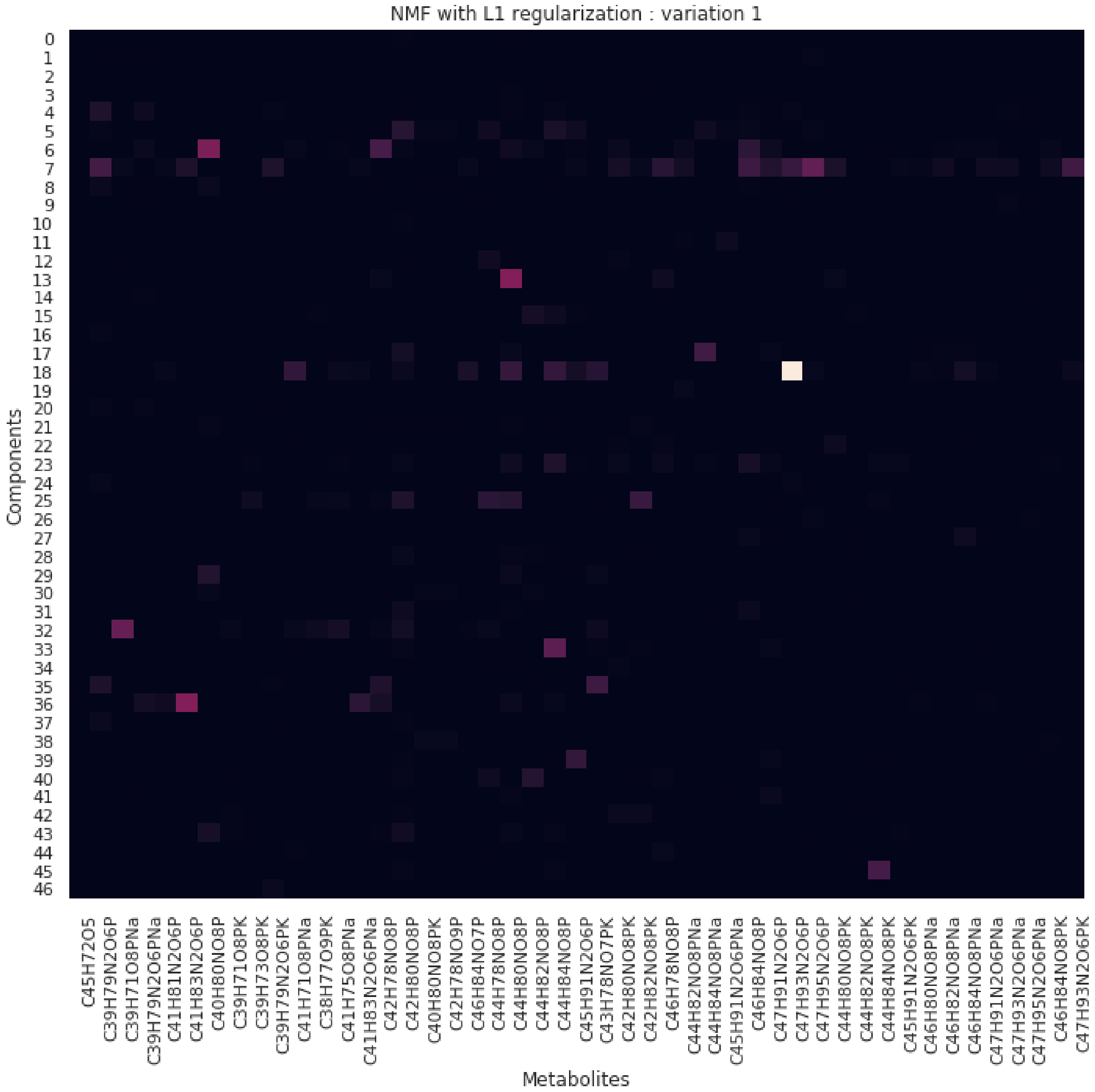
Results of Mass Spec data analysis:

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

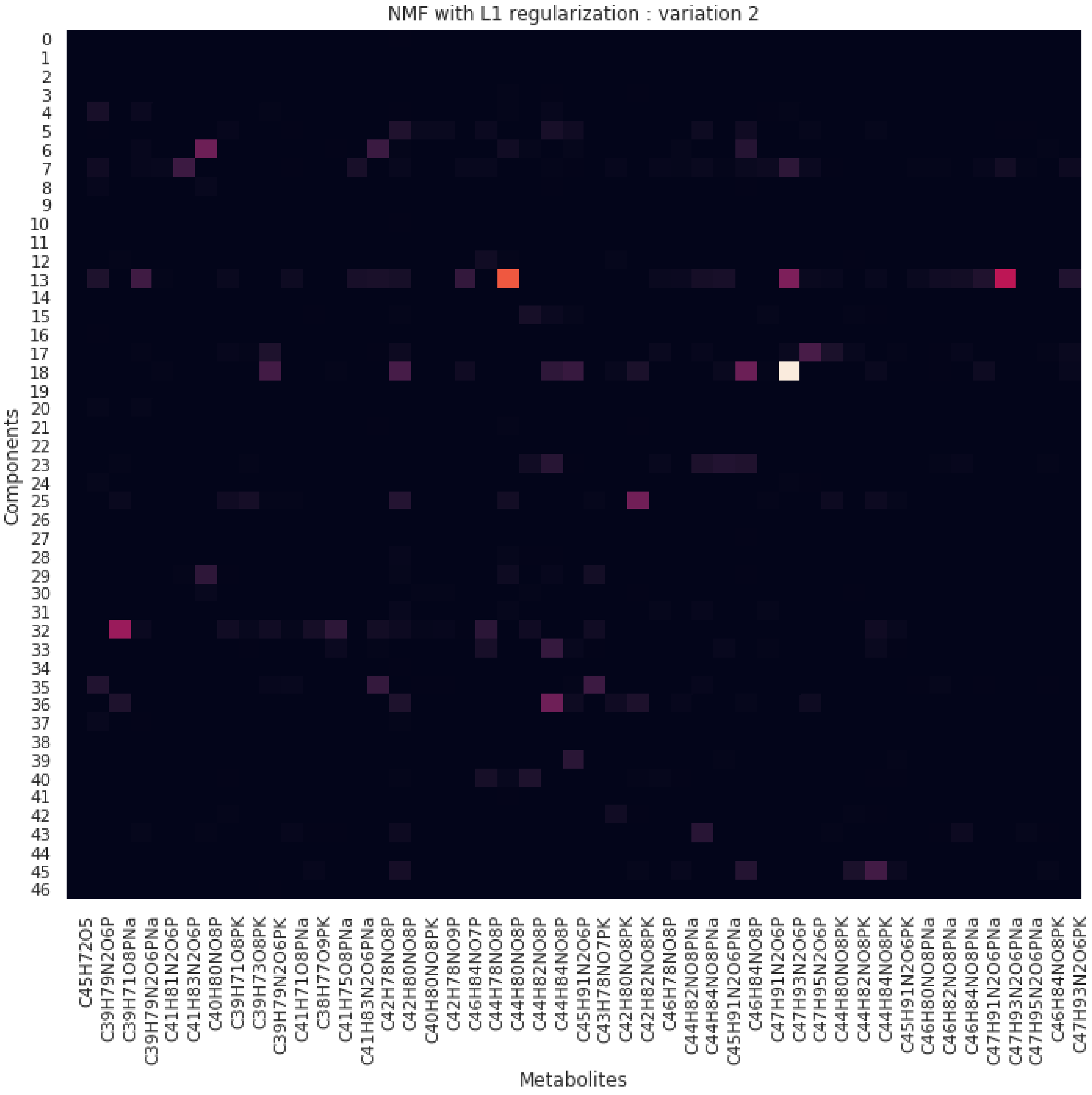


First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu


NMF without L1 regularization
Components
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
Metabolites
C45H72O5
C39H79N2O6P
C39H71O8PNa
C39H79N2O6PNa
C41H81N2O6P
C41H83N2O6P
C40H80NO8P
C39H71O8PK
C39H73O8PK
C39H79N2O6PK
C41H71O8PNa
C38H77O9PK
C41H75O8PNa
C41H83N2O6PNa
C42H78NO8P
C42H80NO8P
C40H80NO8PK
C42H78NO9P
C46H84NO7P
C44H78NO8P
C44H80NO8P
C44H82NO8P
C44H84NO8P
C45H91N2O6P
C43H78NO7PK
C42H80NO8PK
C42H82NO8PK
C46H78NO8P
C44H82NO8PNa
C44H84NO8PNa
C45H91N2O6PNa
C46H84NO8P
C47H91N2O6P
C47H93N2O6P
C47H95N2O6P
C44H80NO8PK
C44H82NO8PK
C44H84NO8PK
C45H91N2O6PK
C46H80NO8PNa
C46H82NO8PNa
C46H84NO8PNa
C47H91N2O6PNa
C47H93N2O6PNa
C47H95N2O6PNa
C46H84NO8PK
C47H93N2O6PK



First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu



Summary of Microscopy images for Mass Spec:

There are two types of data: raw microscopy data and processed microscopy data.
We use only processed Microscopy data for segmentation i.e. to find regions of interest. The Microscopy image is from Fluorescence Microscopy Imaging : refer [7].
For a quick exploratory learning: use ImageJ examples as explained in [7].
The processed Microscopy image has 3 images. We need MxIF to IMS image which contains DAPI channel with details about nuclei. We need a tool that will identify nuclei and segment the image into corresponding cells.

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

Regarding GPU, configuration requirements: Majority of Imaging Mass spec analysis can be done on VM (PSC VMs or on laptop). But for Microscopy imaging we need nvidia based GPU for process intensive steps. Cytokit, for example, is a Docker container and can only be run on Nvidia GPU. However, cell profiler does not need Nvidia GPU but because of the size of the imaging file, we need GPU to process tasks such as segmentation, GUI explorer etc.

Data validation, error checks and accuracy analysis

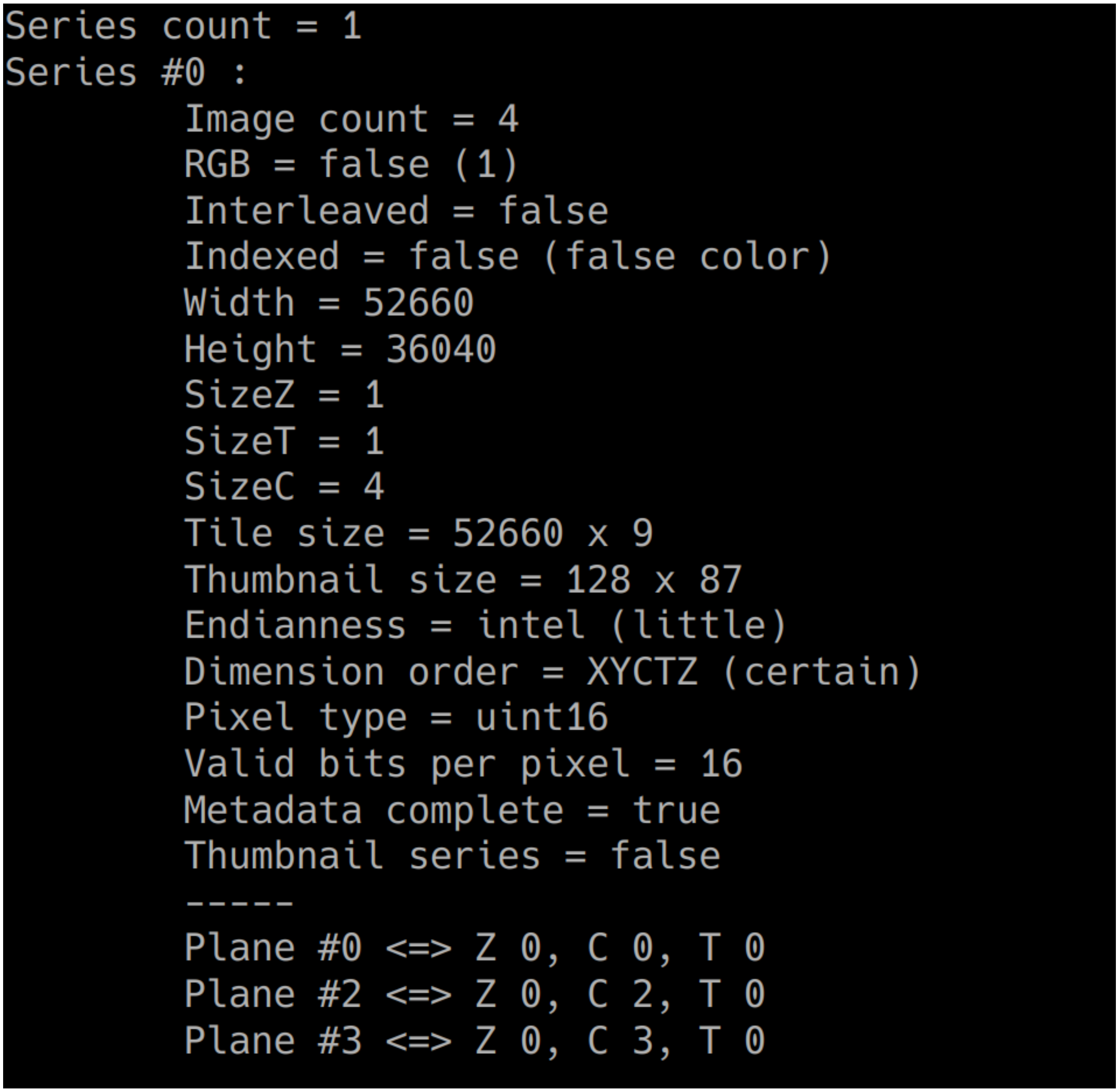
Bftools from Bioformats can be a quick tool for insight into validation of data. However, as per our analysis, bftools or the Cytokit seem to identify the wrong channel (i.e. for DAPI channel) in microscopy images.

- So a preliminary analysis using bftools/showinfo :

```
> bftools/showinf -nopix
hubmap/cytokit/processed_microscopy/VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff
```

Does not detect a 4th channel, which is not abnormal. Each Microscopy image contains XYCTZ values.

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu



```
Series count = 1
Series #0 :
        Image count = 4
        RGB = false (1)
        Interleaved = false
        Indexed = false (false color)
        Width = 52660
        Height = 36040
        SizeZ = 1
        SizeT = 1
        SizeC = 4
        Tile size = 52660 x 9
        Thumbnail size = 128 x 87
        Endianness = intel (little)
        Dimension order = XYCTZ (certain)
        Pixel type = uint16
        Valid bits per pixel = 16
        Metadata complete = true
        Thumbnail series = false
        -----
        Plane #0 <=> Z 0, C 0, T 0
        Plane #2 <=> Z 0, C 2, T 0
        Plane #3 <=> Z 0, C 3, T 0
```

- Depending on analysis of metadata for the Microscopy image, to modify the metadata would be :

```
bftools/tiffcomment -set 'newxml.xml'
../processed_microscopy/VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff
```

- TO check quick metadata details of a cropped Microscopy image, we can crop the image using the bftools, so that cropping will update to metadata can be done using:

```
bftools/bfconvert -crop 20480,20480,2048,2048
hubmap/cytokit/processed_microscopy/VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff
VAN0001-RK-1-21_24-MxIF-mxIF_toIMS_2048_cropped.ome.tiff
```

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

- From an analysis of the Sep 2019 sample data, bftools/other tools seems to identify channel 1 as DAPI. But upon examining the Microscopy images, DAPI channel is indeed 3rd channel. So we need to validate the images to avoid instrument generated/normalized images.

Segmentation of Microscopy images for Mass Spectrometry:

Tools, software available:

There are various tools and methods available to test this. However Cytokit has a segmentation pipeline that uses UNet model to identify segments in a given Microscopy image.
We can also use cell profiler or Deep cell which among other tools to use.
For quick image tools: ImageJ is an informative tool.

Cytokit pipeline and details for September data

Introduction

Cytokit is a

Cytokit Configuration

Pre-requisites:

Cytokit needs an Nvidia GPU, so we use Hive’s GPU or XSede bridges’ GPU.

1. Read access to Hive.psc.edu
2. Read access to XSede and bridges
3. Write access on HIVE and Bridges from respective user’s folder: /hive/users/USERNAME
4. User to have network permissions on GPUs (both on bridges and Hive’s GPU).

Point of Contact for PSC cluster, HIVE / Bridges/ GPUs, access related issues:

For any issues/help required for HIVE clusters, GPU access, XSede bridges: contact help@hubmapconsortium.org and CC to

- Joel Welling <welling@psc.edu>,
- Philip Blood <blood@psc.edu>,
- Matt Ruffalo <mruffalo@cs.cmu.edu>,
- Ziv Bar-Joseph <zivbj@andrew.cmu.edu>

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

Connecting to HIVE and Bridges:

Connect to Hive from terminal:

1. Open terminal and type:

   ```
   ssh your-username@hive.psc.edu
   ```

   and enter password

2. Type following to connect to single GPU here:

   ```
   srun -p GPU --mem=0 --pty /bin/bash
   ```

Connect to bridges from terminal:

1. Open terminal and type:

   ```
   ssh your-username@bridges.psc.edu
   ```

   and enter password

2. Type following to connect to single GPU here:

   ```
   interact -p GPU-small --gres=gpu:p100:2
   ```

Download data, config files required for Cytokit segmentation for running test on Cytokit's Fluorescence images:

1. Create a folder hubmap and hubmap/cytokit/lab/data/ within /hive/users/your-username
2. Now download data required for segmentation:

   ```
   gsutil cp -r gs://cytokit/datasets/cellular-marker ./hubmap/cytokit/lab/data
   ```

3. Download https://github.com/hammerlab/cytokit/blob/master/pub/config/cellular-marker/experiment_dapi.yaml to GPU. You can wget or download locally on your laptop and SFTP to hive.psc.edu or use Filezilla/Winscp and connect to Hive.psc.edu and upload here.
4. This experiment_dapi.yaml must be within following folder structure on GPU:

   /hive/users/your-username/hubmap/cytokit/lab/**repos/cytokit/config/cellular-marker**

Notice that highlighted folder structure, above. You need to create **repos/config/cellular-marker** folder within lab folder that you already have created in previous step.

5. Download the script to run DAPI, PHA examples in Cytokit to understand the configuration and some quick hands-on experience. The script:

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

https://github.com/hammerlab/cytokit/blob/master/pub/analysis/cellular-marker/pipeline_execution.sh Again you can wget, sftp or use filezilla/winscp.

6. You can place the script at /hive/users/your-username/hubmap/cytokit and update DATA_DIR and OUTPUT_DIR according to the locations you have placed the data as in step 3.

Steps to start Cytokit and run pipeline script:

1. $SCRATCH is already created. Open terminal/console.
2. Connect to Hive and run following commands

```
cd $SCRATCH
mkdir singularity
cd singularity
module load singularity
```

3. Since Cytokit runs as a Docker container but HIVE/bridges do not support Docker for security reasons, we will be running it using Singularity. So, we need to pull the Cytokit image from Dockerhub as follows:

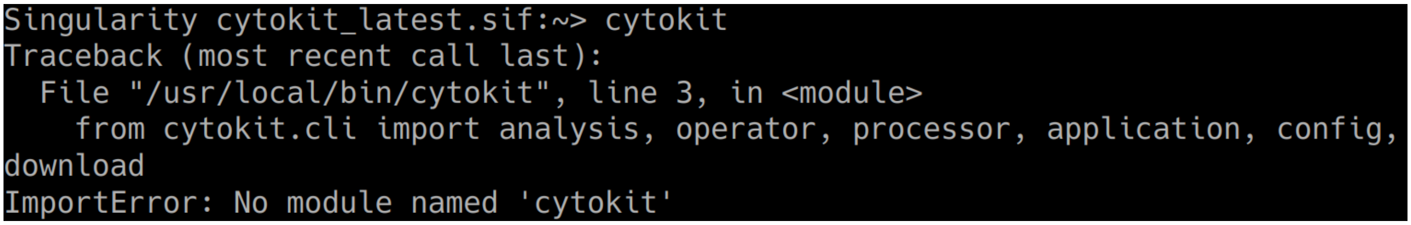
```
singularity pull docker://eczech/cytokit:latest
```

4. Now start the container as follows:

```
singularity shell --nv -B hubmap/cytokit:$SCRATCH/hubmap/cytokit -B
hubmap/cytokit/lab/data:/lab/data singularity/cytokit_latest.sif
```

5. Notice that, once cytokit container starts, you will be able to /lab/data i.e. if you "cd /lab/data" you will be able to see all data. This is the place where .cytokit folder gets created which is a temporary cache, several temporary files are written/deleted.

6. Now, run Cytokit and see that if it loads. Ideally should load automatically. However, if sometimes python3 may or may not globally set up.

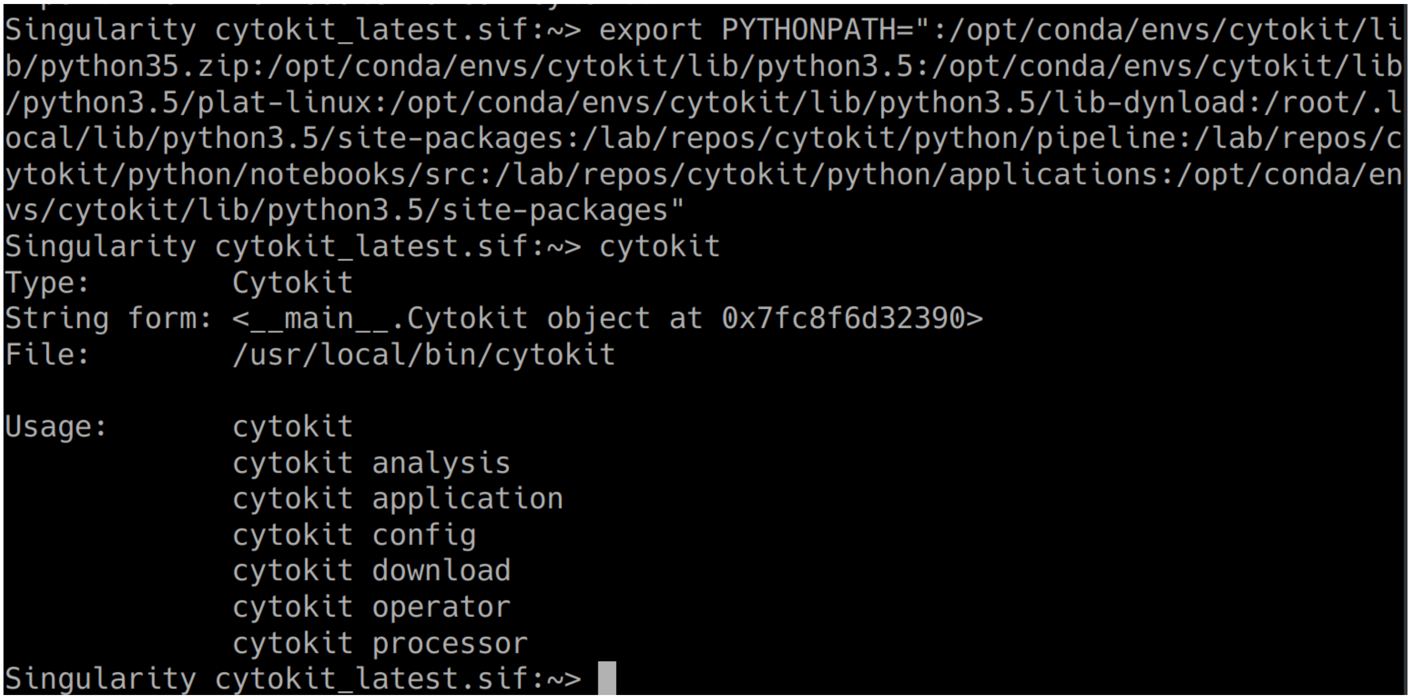


```
Singularity cytokit_latest.sif:~> cytokit
Traceback (most recent call last):
  File "/usr/local/bin/cytokit", line 3, in <module>
    from cytokit.cli import analysis, operator, processor, application, config,
download
ImportError: No module named 'cytokit'
```

7. So if you get error such as "Cytokit: module not found" then simply run following to add python to path variable, as follows:

```
export
PYTHONPATH=":/opt/conda/envs/cytokit/lib/python35.zip:/opt/conda/envs/cytokit/lib/python3.5:/opt/conda/envs/cytokit/lib/python3.5/plat-linux:/opt/conda/envs/cytokit/lib/python3.5/lib-dynload:/root/.local/lib/python3.5/site-packages:/lab/repos/cytokit/python/pipeline:/lab/repos/cytokit/python/notebooks/src:/lab/repos/cytokit/python/applications:/opt/conda/envs/cytokit/lib/python3.5/site-packages"
```

8. Now, type Cytokit at command prompt and it should look like this:



```
Singularity cytokit_latest.sif:~> export PYTHONPATH=":/opt/conda/envs/cytokit/lib/python35.zip:/opt/conda/envs/cytokit/lib/python3.5:/opt/conda/envs/cytokit/lib/python3.5/plat-linux:/opt/conda/envs/cytokit/lib/python3.5/lib-dynload:/root/.local/lib/python3.5/site-packages:/lab/repos/cytokit/python/pipeline:/lab/repos/cytokit/python/notebooks/src:/lab/repos/cytokit/python/applications:/opt/conda/envs/cytokit/lib/python3.5/site-packages"
Singularity cytokit_latest.sif:~> cytokit
Type:        Cytokit
String form: <__main__.Cytokit object at 0x7fc8f6d32390>
File:        /usr/local/bin/cytokit

Usage:       cytokit
             cytokit analysis
             cytokit application
             cytokit config
             cytokit download
             cytokit operator
             cytokit processor
Singularity cytokit_latest.sif:~>
```

9. By default, singularity container starts at /hive/users/your-username.

10. To run Cytokit pipeline on Cytokit's provided DAPI/PHA images under cellular-marker folder. So navigate to hubmap/cytokit and execute pipeline_execution.sh as follows:

```
sh pipeline_execution.sh
```

11. Typically, this is the step that seems like you are struck. But 99% of the time, all errors that you now encounter either depend on write access you have to folder that is mapped to /lab/data or that your configuration i.e. experiment.dapi.yaml does not necessarily work or pipeline_execution.sh highly likely did not have correct DATA_DIR or Output_DIR values.

12. By default all HUBMAP data is available under: /hive/hubmap/lz/ . Specifically, Vanderbilt's Microscopy images for Mass Spec are in /hive/hubmap/lz/Vanderbilt TMC.

Steps to modify experiment.yaml and run cytokit commands for Vanderbilt's Fluorescence images from September 2019:

NOTE: No need of script for this. We just simply run cytokit processor and analysis commands here.

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

We just need to provide correct file/folder paths for experiment.yaml, data folder, output folders.

1. Edit the VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff's metadata xml as follows
   bftools/tiffcomment -set 'newxml.xml'
   ../processed_microscopy/VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff
   Download newxml.xml from
   https://drive.google.com/open?id=1bLI_0aa8G9ZD1KYisqyBQ2Z9S9C6uLie

2. bftools/tiffcomment ../processed_microscopy/VAN0001-RK-1-21_24-MxIF-mxIF_toIMS.ome.tiff

3. Download experiment.yaml
   https://drive.google.com/open?id=1R5jHQh9O1Pp27mfc0y51kISIyCdsvkTT and copy to
   /hive/users/your-username/hubmap/cytokit/

4. For Cytokit processor step, run following command:

   ```
   cytokit processor run_all --config-path=hubmap/cytokit/experiment.yaml
   --data-dir=/lab/data/cytokit/cellular-marker/van_ms_dapi
   --output-dir=/lab/data/cytokit/cellular-marker/van_ms_dapi/output
   ```

5. Once the above step is successful, run Cytokit analysis step as follows:

   ```
   cytokit analysis run_all --config-path=hubmap/cytokit/experiment.yaml
   --output-dir=/lab/data/cytokit/cellular-marker/van_ms_dapi/output
   ```

6.

Why do we need to generate tiles?

bfconvert docs say that "%m" is the overall tile index that does not necessarily suggest tiles are generated in which sequence: snake or continuous vertical etc. Whether the way that bfconvert indexes tiles would be the same as what Cytokit is expecting, in the Cytokit config we can either set tiling mode to "snake" or something else (have to check the code to remember the other mode name).
"snake" expects the tile indices to be read left --> right in the first row, and then right --> left in the second row, then left --> right in the third row, etc. it's just to do with the way the microscope moves when it's acquiring images. Something like

```
# 400 tiles
# 20 x 20
# "Snake" pattern, e.g.:
#   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
#  40  39  38  37  36  35  34  33  32  31  30  29  28  27  26  25  24  23  22  21
#  41  42  42  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
```

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

```
#    etc
```

so in the Cytokit code there is a way to convert from X,Y coordinates to snake tiling indices, and back and I used this to rename the files generated by bfconvert so they had the tile indices in instead of the X, Y coordinates.

Steps to run entire cytokit, including tile generation and starting container:

1. Fluorescence Microscopy image that is provided by Vanderbilt a single image. For Cytokit, we need to generate tiles to provide as input. So we use this script to generate tiles
2. Since tiles generated by bfconvert command contain numbering starting at 1, however, Cytokit does skips such image names. And series, plane and channel number also must start from 1, and hence every number must be incremented by 1. For that, we use this script to rename tiles that are generated.
3. To start singularity's Cytokit container, this script starts the container.
4. To run Cytokit's processor command, this script does the work.

Output of Cytokit Pipeline for Segmentation task:

Note: results included here were run by Maria Keays on bridges.psc.edu.

1. After segmentation, the tile 170 with segmented nuclei would look like https://drive.google.com/open?id=1W-nIFW0BY9tBNXmU72-I27ZkxLXJ_4jQ.

2. And tile 170 would look like this. https://drive.google.com/open?id=13ELLUzcb_r0ZNF1urfzxAdxdD4DCOB45

3. The tile 85's channels would look like this: https://drive.google.com/open?id=12EtoTvFDgDOa8P7xwiPMCmInOE-d2f6u

4. The output of Cytokit processor and analysis steps are as follows, as shared by Maria Keays:
   - output/cytometry directory contains segmentation results:
     - tile directory:
       - one TIFF per tile, each with 4 channels: 1) cells segmentation mask, 2) nuclei segmentation mask, 3) cell boundaries, 4) nuclei boundaries.
     - statistics directory:
       - one CSV per tile, with cytometric measurements, one row per cell.
     - data.csv: aggregated cytometric measurements from above CSVs, one row per cell. data.fcs has the same data but in FCS format.
   - output/processor directory contains pre-processing results:
     - tile directory:
       - one TIFF per tile, each with 4 channels.

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

        - execution directory:
            - JSON files with parameters of each Cytokit run.
        - data.json: results of best focus plane selection, irrelevant here as there’s only one plane.

Troubleshooting

A significant amount of effort to formalize steps to start cytokit successfully was required. Most likely the above steps will ensure cytokit starts and script runs without any hiccups. However, if you still do encounter any errors, this section attempts to address, likely all of them.

1. **Error message:** FATAL: container creation failed.
   When does this occur: when you attempt to run singularity shell --nv -B command from step 4 in Steps to start Cytokit: section.
   A very common, likely error, since not many users actively use or do actively use and sometimes, this error can occur. Some incorrect deletion by a GPU maintenance scripts/weekly updates etc could often lead to this error. The error would be like:
   FATAL: container creation failed: failed to resolved session directory /mnt/singularity/mnt/session: lstat /mnt/singularity: no such file or directory
   Solution: Typically, at hive.psc.edu, if you navigate to /mnt/singularity, this folder should exist. The same folder structure must exist once you login to GPU from HIVE or bridges GPUs. So in either of GPU, you will not have access to /mnt/singularity as you are NOT an admin to HIVE or bridges. So these folders should be available, ideally. If you dont find /mnt/singularity in either of GPUs, then follow these steps to inform PSC team.

2. Sometimes, GPU is not connected successfully, due to some updates/reconfiguration. Then follow these steps to inform PSC team.

3. **Error :** DATA_DIR or BASE_CONF directory not found error when running pipeline_execution.sh
   When does this occur: when DATA_DIR is not correctly mapped.
   Solution: DATA_DIR must be something like
   DATA_DIR=/lab/data/cytokit/cellular-marker/$EXP_NAME
   And BASE_CONF must be something like:
   BASE_CONF=lab/repos/cytokit/config/cellular-marker/experiment_$EXP_TYPE.yaml
4. **Error using Cytokit’s cellular marker data (downloaded from gsutils (not HUBMAP data)):**
   Config file not found in variant: /lab/data/cytokit/cellular-marker/
   20180614_D22_RepA_Tcell_CD4-CD8-DAPI_5by5/output/config/v00/experiment.yaml
   Or
   /lab/data/cytokit/cellular-marker/
   20180614_D22_RepA_Tcell_CD4-CD8-DAPI_5by5/output/config/v01/experiment.yaml
   When does this occur: In pipeline_execution.sh that you run in step 9 from Steps to start Cytokit: , in https://github.com/hammerlab/cytokit/blob/1d96e437d16b9f12bbbf02198a975f81904cfdbe/pub/analysis/cellular-marker/pipeline_execution.sh#L24 step, for each variant, a config/v0X/experiment.yaml will be generated. So this config generation has not been successful. So there are two things to note

here: If you are using Cytokit data that you downloaded in Step 2 from [Download data section](), then you will need to inspect why config generation is not successful. Sometimes, it is the resources, memory etc that could cause.

- ➢ One solution is to just use one of the Experiments here and see what is happening https://github.com/hammerlab/cytokit/blob/1d96e437d16b9f12bbbf02198a975f81904cfdbe/pub/analysis/cellular-marker/pipeline_execution.sh#L4
- ➢ Another solution is to modify and run pipeline_execution.sh only for DAPI and PHA files separately: i.e. keep only Experiments folder names such as follows:

```
EXPERIMENTS="
20180614_D22_RepA_Tcell_CD4-CD8-DAPI_5by5;dapi;35
20180614_D22_RepB_Tcell_CD4-CD8-DAPI_5by5;dapi;35
20180614_D23_RepA_Tcell_CD4-CD8-DAPI_5by5;dapi;35
20180614_D23_RepB_Tcell_CD4-CD8-DAPI_5by5;dapi;33
“
```

Or only for PHA:

```
EXPERIMENTS="
20181116-d40-r1-20x-5by5;pha;25
20181116-d40-r2-20x-5by5;pha;25
20181116-d41-r1-20x-5by5;pha;25
20181116-d41-r2-20x-5by5;pha;25
"
```

5. Caution: When you run Cytokit pipeline script on Cytokit's provided cellular marker data, you do NOT have to modify experiment_XXX.yaml at all.The purpose of the config files https://github.com/hammerlab/cytokit/blob/master/pub/analysis/cellular-marker/pipeline_execution.sh is that it should run as is for Cytokit cellular marker data.

6. For stitching images, refer: https://github.com/hammerlab/cytokit/issues/16 [github.com])

7. Another error, possibly due to ResourceExhaustionError can occur.
Error: ResourceExhaustionError

```
Exception: ('Long error message', <class
'tensorflow.python.framework.errors_impl.ResourceExhaustedError'>, 'OOM when allocating tensor
with shape[1,192,1808,2640]
```

When does this error occur: This is because per-user memory on GPU RAM is limited as it could have been allocated by one user and then another user's request to allocate memory could be denied.
Solution: use bridges.psc.edu's gpu from step
Map /hive/hubmap/lz/sample/MKeays_Share/Cytokit_testing/cytokit_cache_data/lab/data as /lab/data when launching singularity cytokit container i.e.

```
/hive/hubmap/lz/sample/MKeays_Share/Cytokit_testing/cytokit_cache_data/lab/data
```

8.

Cell Profiler for Microscopy image

Introduction

In short, Cell Profiler implements segmentation algorithm. Cytokit, in short, uses Cell Profiler for segmentation task. To get an intuition of each step in cell profiler, install Cell Profiler and use Explorer to run on example.

Configuration

Pre-requisites:

1. You need Docker

Steps to Install:

1. Pull Cell profiler docker container

   ```
   docker pull cellprofiler/cellprofiler
   ```

2. Start Docker container:

   ```
   docker run -it -e DISPLAY=$DISPLAY -v /tmp/.X11-unix:/tmp/.X11-unix:ro cellprofiler/cellprofiler:3.1.9 ""
   ```

3.

Troubleshooting

December Data release 2019:

Changes from September 2019

What next ?

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu

Outstanding questions and work-in-progress:

Questions

Work in Progress

References:

1. Mass Spectrometer https://www.khanacademy.org/science/in-in-class-12th-physics-india/moving-charges-and-magnetism/in-in-magnets-and-magnetic-force/v/mass-spectrometer
2. MALDI MS : https://pubs.rsc.org/image/article/2018/RA/c8ra01574k/c8ra01574k-f4_hi-res.gif
3. MALDI :https://www.youtube.com/watch?v=8R1Oyqx5KfE
4. How to prepare sample for MALDI: https://www.youtube.com/watch?v=-PlrQVxtVE0
5. Vanderbilt IMS workflow : https://drive.google.com/drive/u/0/folders/1k0BeW0ijCL70JFjMr9U1P3pQqD0Ttd0o
6. Vanderbilt's webinar on MALDI : https://drive.google.com/open?id=10yOUGncEVudG4IcCelGqwf_USb8wFGZy
7. Details of Vanderbilt's sample data from September: https://docs.google.com/document/d/12WhQK5Y5cHju5C2yicw8TQ5V5vOma2ZHJtLEI_vj0Jk/edit
8. Fluorescence Microscopy Imaging :https://petebankhead.gitbooks.io/imagej-intro/content/chapters/thresholding/thresholding.html
9. Pipeline execution script: https://github.com/hammerlab/cytokit/blob/1d96e437d16b9f12bbbf02198a975f81904cfdbe/pub/analysis/cellular-marker/pipeline_execution.sh
10. Metabolomics database: https://www.metabolomicsworkbench.org/data/mb_mass_form.php
11. Mass Spec Coloc Metaspace: https://github.com/metaspace2020/coloc
12. Mass Spec Coloc Metaspace No learning pipeline: https://github.com/metaspace2020/coloc/tree/master/measures/NoLearning
13. Script to generate tiles from Microscopy image: https://drive.google.com/open?id=1Ke1qykeMCy3wOICherfHfdIncHEVjgFo
14. Script to rename tiles to expected format for Cytokit: https://drive.google.com/open?id=1WBQBLvXKjKmzHWyzxG-y9HeOxJieYQYQ
15. Script to start Singularity container for Cytokit: https://drive.google.com/open?id=1GLpEfoIDLog5Qbjj0TBLeX4rCKOOs72u
16. Script to start Cytokit processor step: https://drive.google.com/open?id=1-1MrMmOE5xxbinNaa6pKEJ2HWdtHd1on

First draft: by Sushma Anand Akoju, email: sakoju@andrew.cmu.edu