Pathology Whole Slide Image De-Identification

MIDI Workshop
Day 2
Panelists

Adam Taylor
Sage Bionetworks

David Gutman
Emory University

Tom Bisson
Charité Universitätsmedizin Berlin
Format

- Introduction
  - Digital pathology and whole slide imaging in the context of de-identification
  - De-identification validation for WSI and multiplexed tissue imaging in HTAN
- Presentations on recent work
  - David Gutman
  - Tom Bisson
- Panel discussion
Digital pathology and whole slide imaging

("digital pathology" OR "whole slide imaging") AND "de-identification"
De-identification considerations for WSI

Filename: TCGA-V1-A8MU-01Z-00-DX1.2C9CED13-5C2C-4FFB-AB0B-3904BAA4FEFF.svs (864,7 MB)

Slide label

Burnt in text

Image headers/tags

TIFF public tags

Assay/study metadata

DateTime 306 (0x132) ASCII: 2020:01:28 14:32:26
Format and metadata diversity

Bioformats supports 162 formats with varying degrees of metadata readability.
Multiplexed tissue imaging

We are in an expansionist phase in the development of advanced multiplexed tissue imaging modalities. We should anticipate their adoption in research and clinical environments.

Increasingly larger, higher dimension and higher resolution data
Sage Bionetworks

Accelerate biomedical discoveries by improving methods for scientific collaboration and communication
Human Tumor Atlas Network

A National Cancer Institute funded Cancer Moonshot Initiative

Project goal(s): Construct 3D atlases of the dynamic cellular, morphological and molecular features of human cancers as they evolve from precancerous lesions to advanced disease.

Level 1 and 2 Sequencing ➪ NIH CRDC CDS authenticated through dbGaP available through Seven Bridges Genomics Cloud

Level 3 and 4 Sequencing ➪ Available through Synapse, and through Google BigQuery at ISB-CGC (June 2022)

Imaging ➪ CDS and Imaging Data Commons
Data in the context of additional metadata

Metadata such as ‘Sectioning days to index’ where sectioning is proximal to imaging provides an attack vector to reconstruct participant date of birth
Data in the context of additional metadata

This effect is multiplied in the case of longitudinal studies
Validation of de-identification in HTAN

1. Proposed assay 1
   - Data survey
   - De-ID plan
   - Generated data
   - Collected metadata

2. De-identification
   - Oops! You found some PHI/PII
   - Center determination
   - Potential PHI
   - Affirmation
   - False positive
   - De-ID validation
   - S3/GCS/Synapse
   - Data Curator App
   - Validation/binning

3. You tried to submit PHI

4. Center
   - DCC
   - Validated
   - Excluded

5. Systems:
   - Portal
   - BigQuery
   - CRDC
Validation of de-identification in HTAN
FAIR data requires communication at scale

Data contributors

- We have a dataset with some files. How should we submit it?
- We want to know if these data are useful for our problem.

Data coordinators

- How do we know what problem is relevant for your data? Describe data well.
- How should we describe our data? We need this out soon!
- Contextualize your data. That way it’d be interoperable too!

Data users

- This dataset is useful! How can we find more like it?
- We want to know if these data are useful for our problem.
- Say WHAT now?!