Session III: Data Sharing

co-PI: Ethan Cerami, X. Shirley Liu

presented by James Lindsay

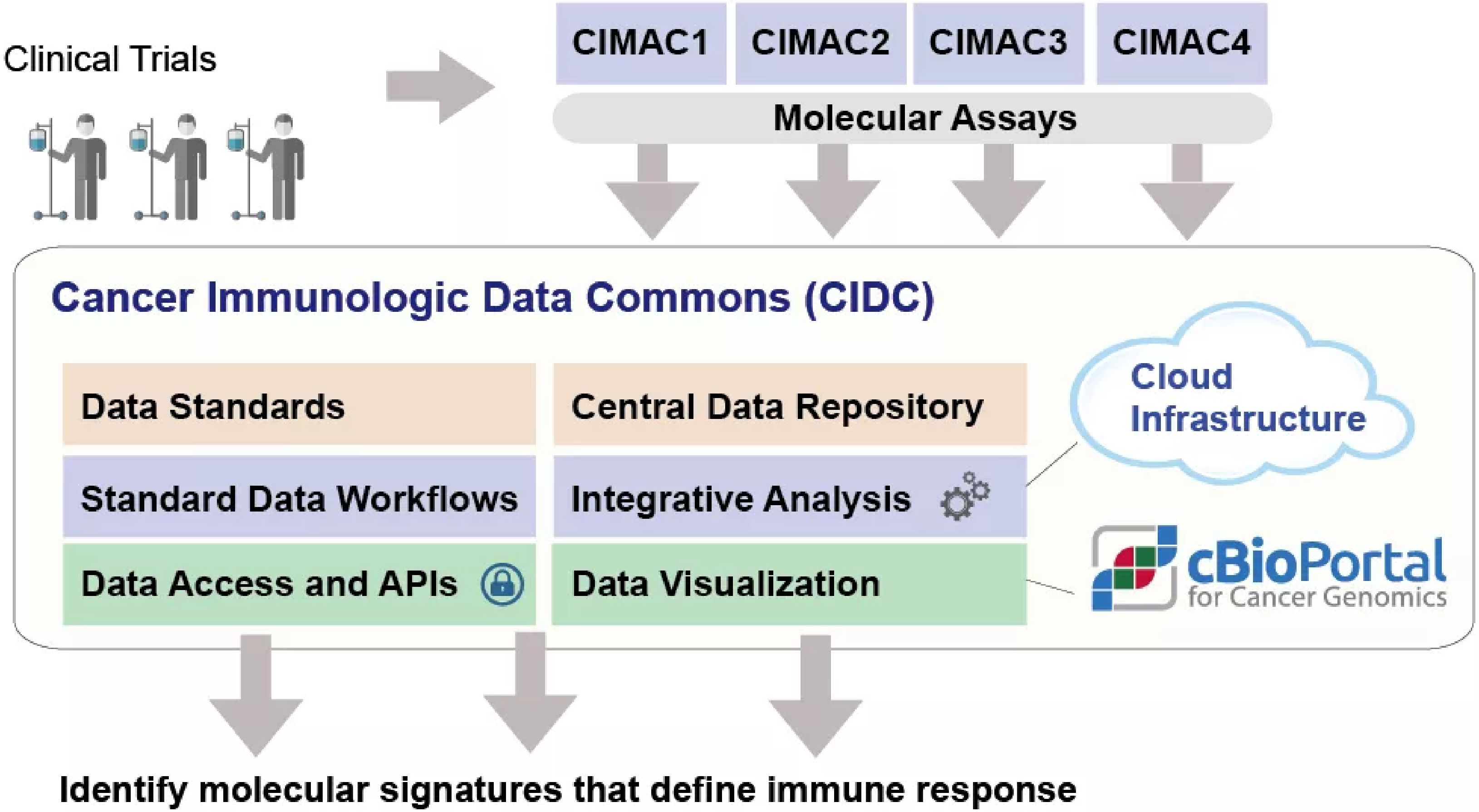
What is a data commons

Scientific data commons enable **wide sharing of information**

-- data, software, and methods.

CIMAC-CIDC: Why do some patients respond to immunotherapy and others do not?

Overview of the CIMACs/CIDC Immunotherapy Network



Clinical Trials

CIMAC1

CIMAC2

CIMAC3

CIMAC4

Molecular Assays

Cancer Immunologic Data Commons (CIDC)

Data Standards

Central Data Repository

Standard Data Workflows

Integrative Analysis

Data Access and APIs

Data Visualization

Cloud Infrastructure

cBioPortal for Cancer Genomics

Identify molecular signatures that define immune response

Thoughts on sharing

Availability

Get data and tools into the hands of researchers fast, remove roadblocks

Community

Build a community around our software and bioinformatics tools (emulate TCGA model)

Innovation

Focus on software and visualization unique to immune biomarker space

Genomics data

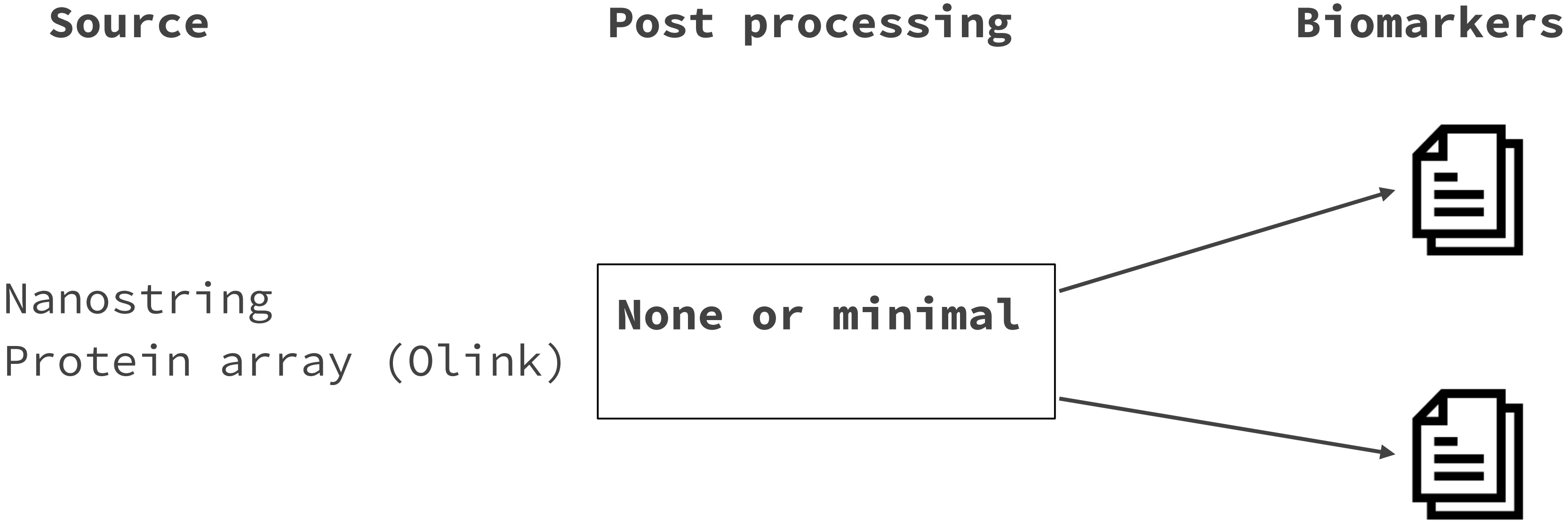
Level 1: High priority data types which will likely be generated in year 1

- Whole exome DNA-seq
- Bulk RNA-seq / Nanostring
- CyTOF
- Singleplex IHC
- Protein array (Olink)
- ** Multiplex IF
- ** TCR sequencing

Future: Other genomics data types under consideration

- Multiplexed Ion Beam Imaging (MIBI)
- Single cell RNA/TCR/BCR…
- 16S sequencing (microbiome)
- RNA-FISH
- HiDim Flow cytometry
- etc...

Genomic data harmonization: The easy



Source

Post processing

Biomarkers

Nanostring

Protein array (Olink)

None or minimal

Genomic data harmonization: The OK

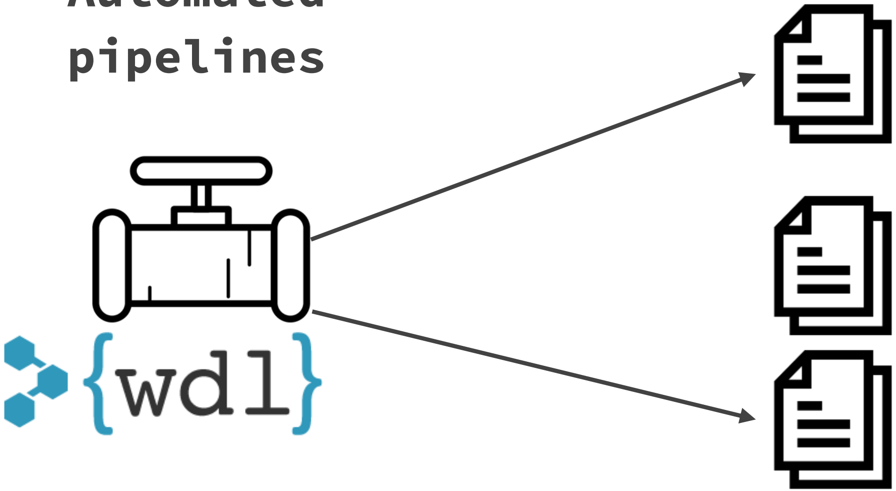
Source

Post processing

Biomarkers

Whole exome DNA-seq
Bulk RNA-seq
TCR sequencing
Single cell RNA/TCR/BCR…
16S sequencing
**CyTOF

Automated pipelines



Genomic data harmonization: The challenging

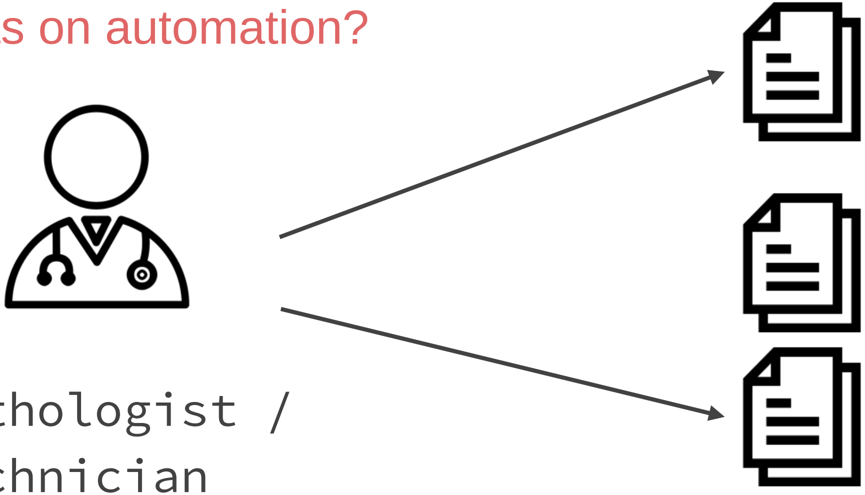
Source

Post processing

Biomarkers

Singleplex IHC
Multiplex IF
Multiplexed Ion Beam Imaging (MIBI)
RNA-FISH
HiDim Flow cytometry
etc...

Thoughts on automation?



Pathologist / Technician

Clinical data

Long-term vision

Defined by NCI, using existing standards such as **CDISC**

Problem

Participating trials use many difference standards and systems

Current status

Develop ad hoc model using ETCTN #10021 replace this ASAP

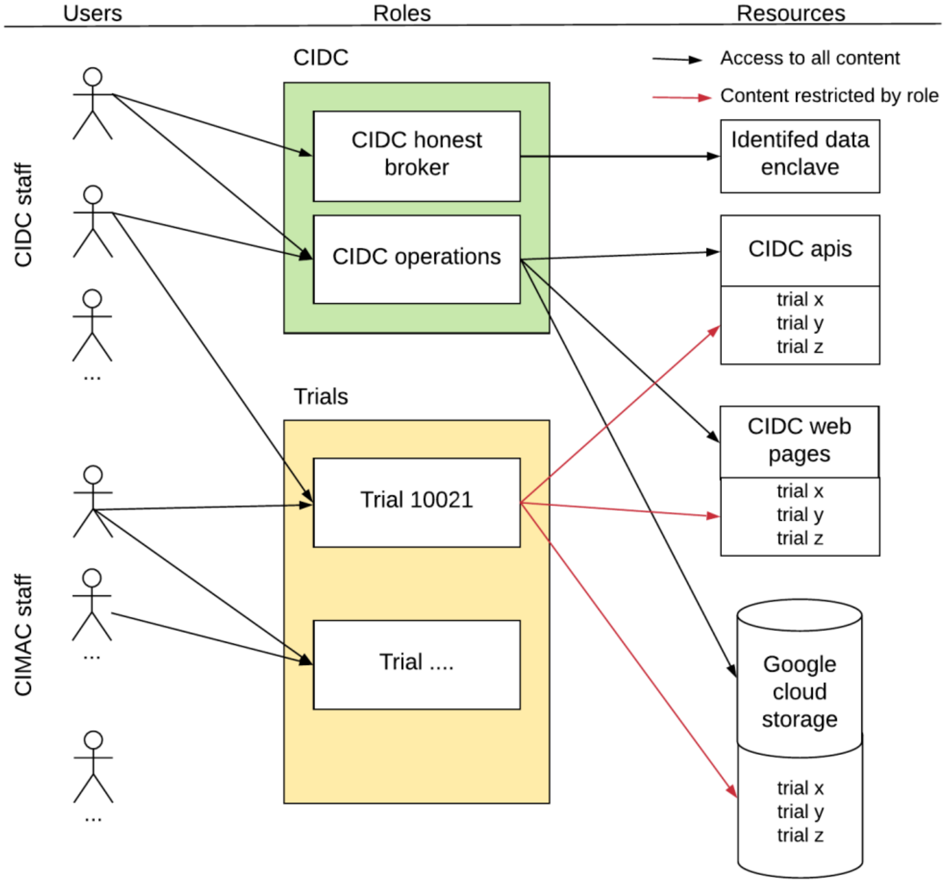
Sharing

Note on security

All content will be secured using industry standard practices

System will be FISMA moderate compliant (eventually)

Role based access control on all resources



FAIR data

Findable

Accessible

Interoperable

Reusable

CIDC-CIMAC network is committed to these guiding principles

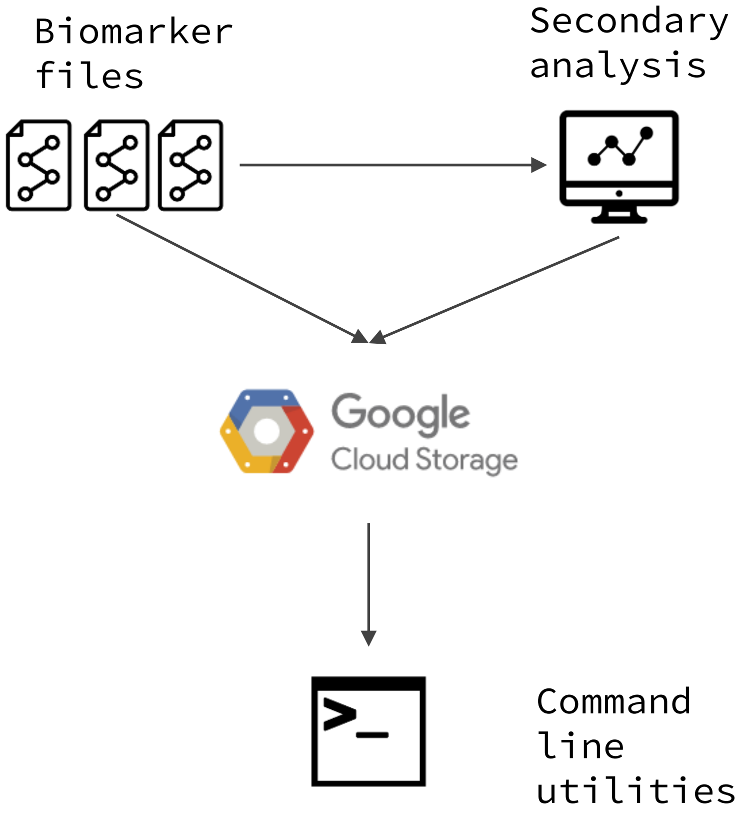
Data sharing modalities

1. Primary and derived files
2. Standardized biomarker calls via API
3. Integration with FireCloud [bring compute to data]
4. Data science interfaces

Primary and derived files

All files generated by bioinformatics also stored in google cloud

There will likely be tiered access to files similar to TCGA



Biomarker files

Secondary analysis

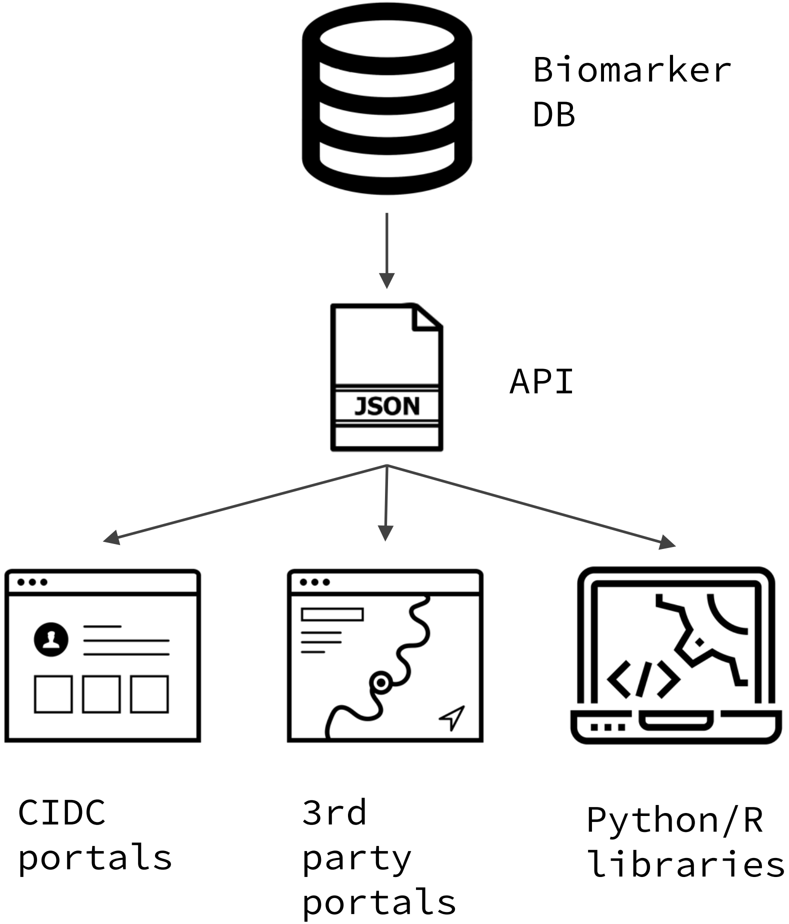
Google Cloud Storage

Command line utilities

Standardized biomarker API

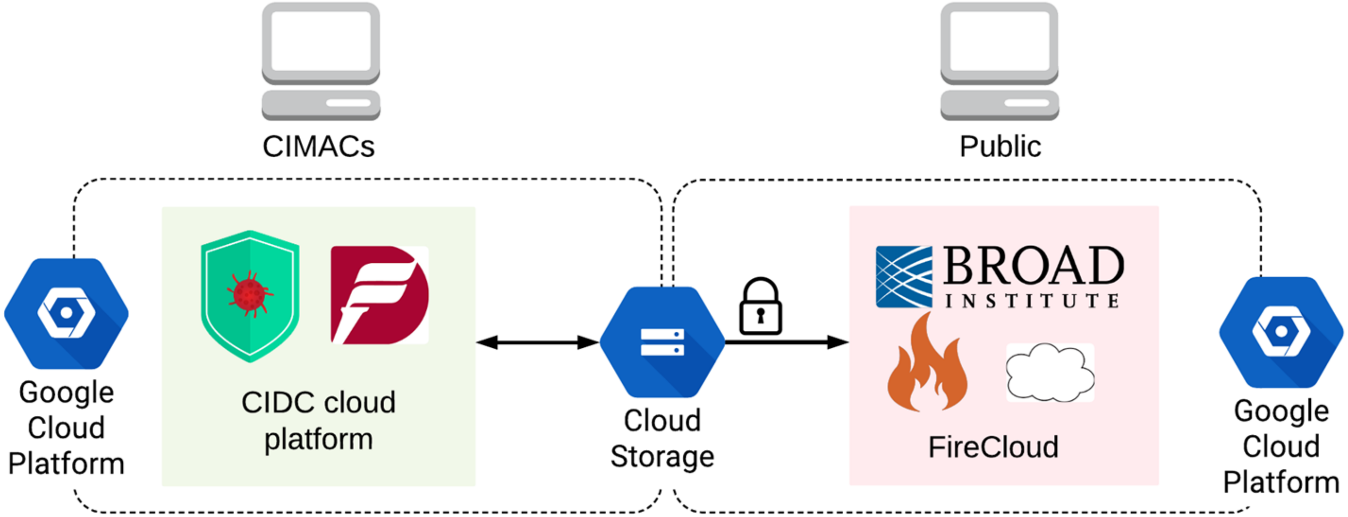
Biomarker + clinical data are stored in a database

Programmatic access to biomarkers via web API



FireCloud

“Bring compute to the data”



Data science interfaces

CIDC data browser
Find data of interest

Browse results of standardized analysis

The ecosystem

cBioPortal
FOR CANCER GENOMICS

Dash

Shiny
by RStudio

jupyterhub

Thoughts on sharing

Availability

Get data and tools into the hands of researchers fast, remove roadblocks

Community

Build a community around our software and bioinformatics tools (emulate TCGA model)

Innovation

Focus on software and visualization unique to immune biomarker space