Seven Bridges Cancer Genomics Cloud
NYU

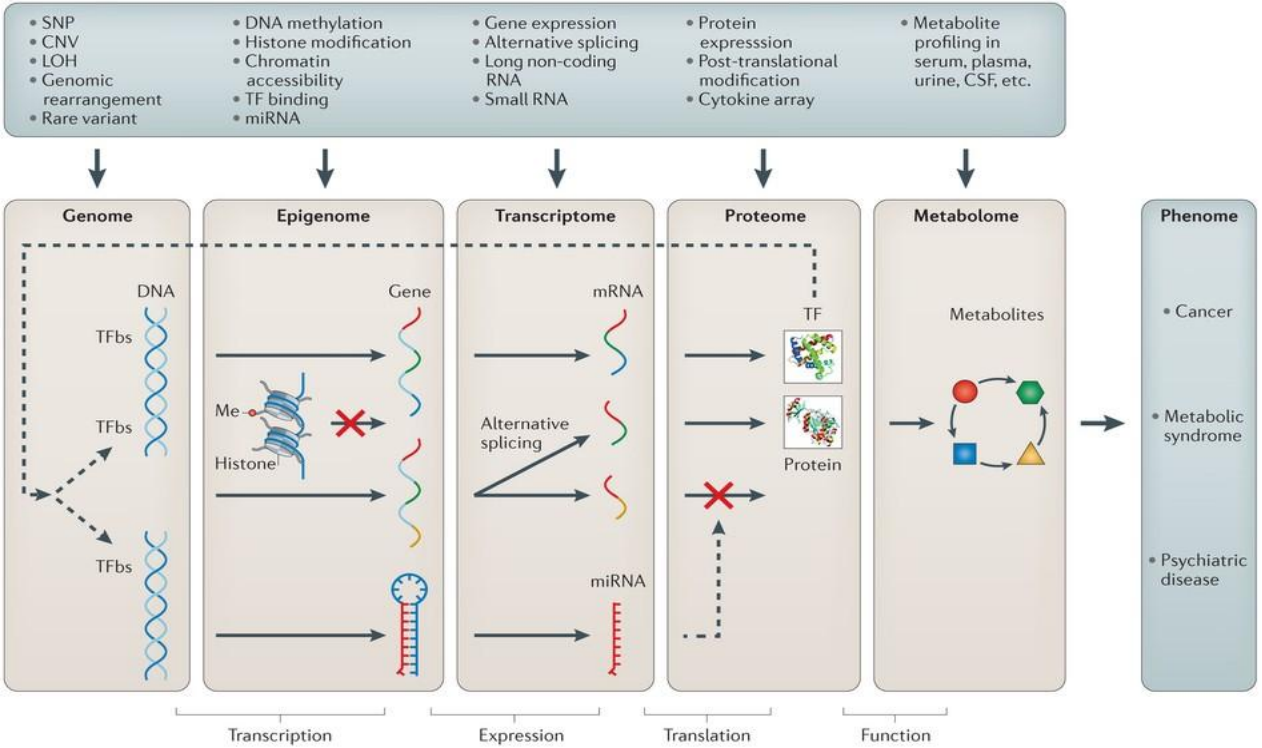
Integrating multi-omic data on the cloud yields insights into retrotransposon activity in cancer

September 13, 2019

Manisha Ray, Ph.D (Seven Bridges)
Wilson McKerrow, Ph.D. (NYU)

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Multi-omic data is critical for cancer research



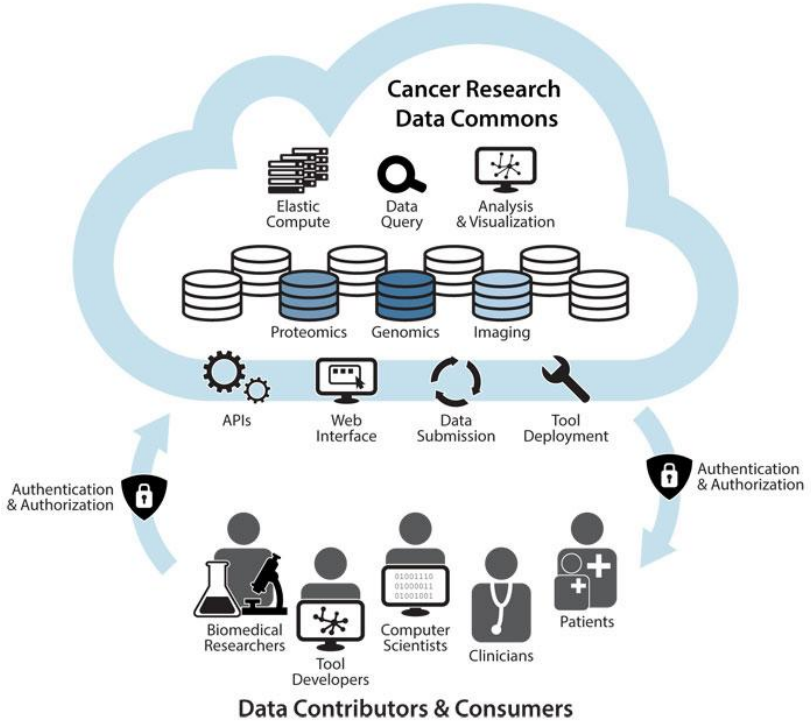
Ritchie et al., Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome to the phenome. (2015)

Comprehensively understanding the full picture of a research question requires examining multiple modalities

© 2019 Seven Bridges

The Seven Bridges Cancer Genomics Cloud

CANCER GENOMICS CLOUD
SEVEN BRIDGES



A Cloud Resource within the NCI Cancer Research Data Commons for secure storage, sharing & analysis of petabytes of public, multi-omic cancer datasets

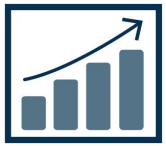
NATIONAL CANCER INSTITUTE

The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and Task Order No. 17X146 under Contract No. HHSN261201500003I.

The Seven Bridges Cancer Genomics Cloud

- A stable, secure, and highly customizable cloud storage and computing platform
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines
- Register for free today at **cancergenomicscloud.org**

Easy data management

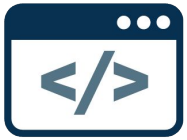


Scalable computation

Optimized bioinformatics algorithms

Secure collaboration

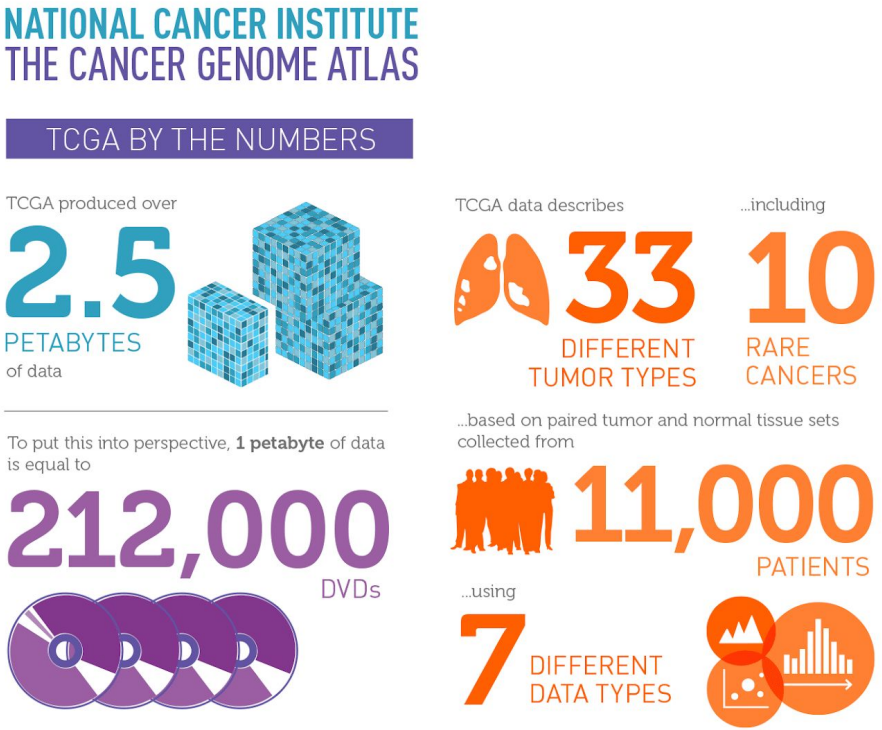
Flexible & fully reproducible methods



Extensible and developer-friendly platform

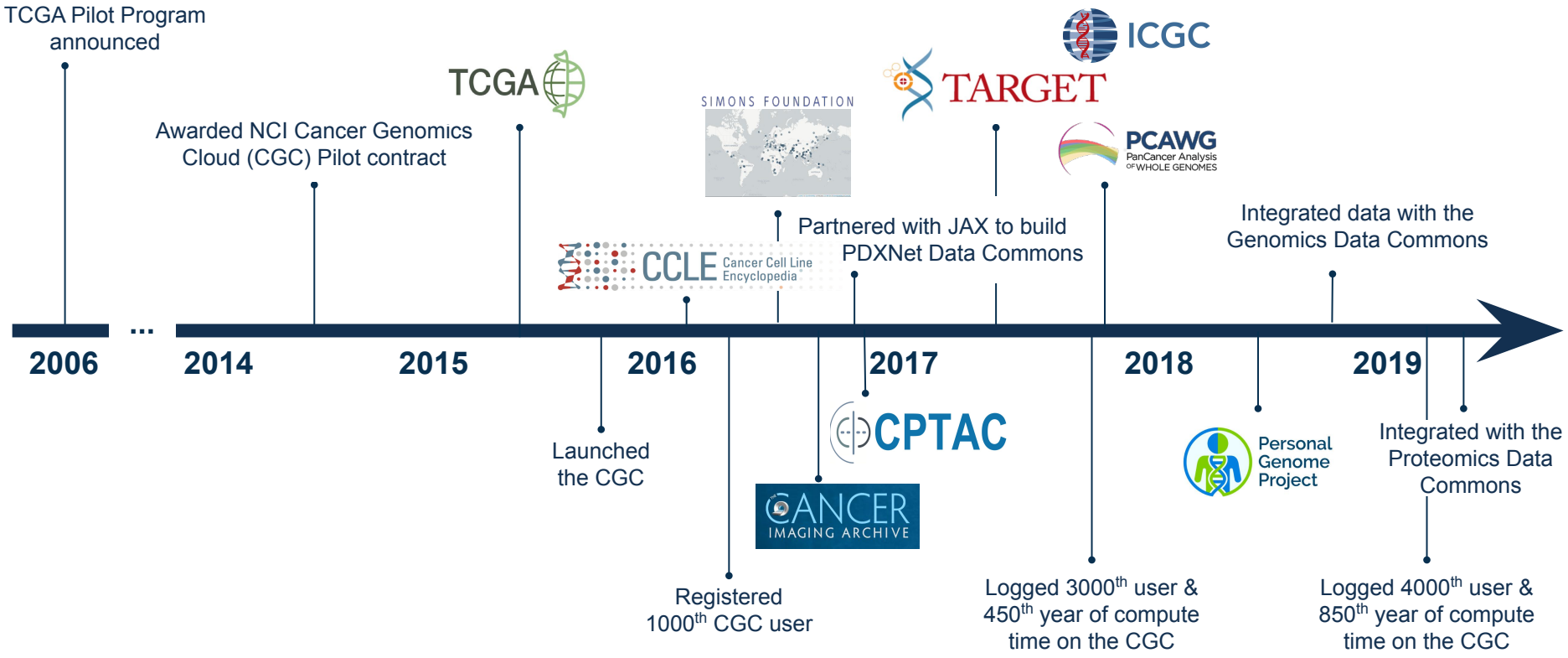
© 2019 Seven Bridges

Need for Cloud Resources was motivated by the growth of TCGA



www.cancer.gov/ccg

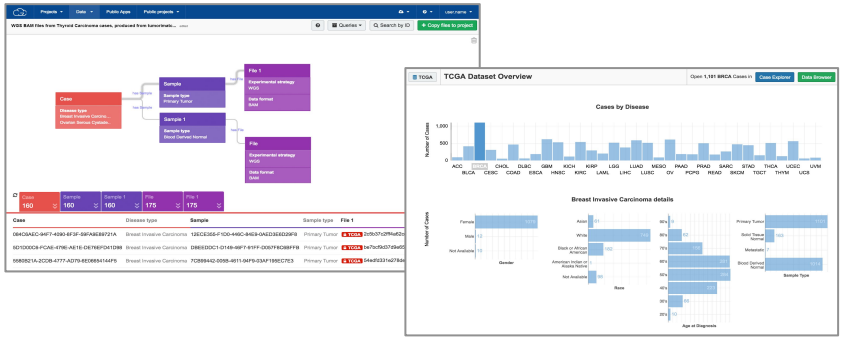
Growth of the Cancer Genomics Cloud Ecosystem

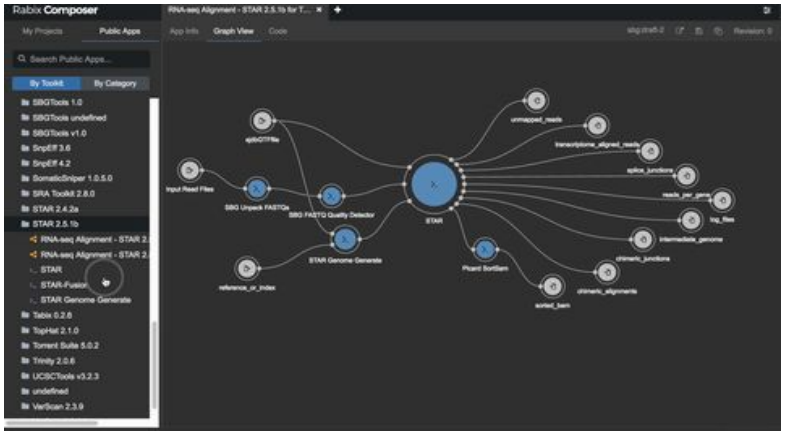


The Seven Bridges Cancer Genomics Cloud (CGC)

- Access **3+ PB** of multi-omic public data through interactive query tools & APIs.
- Upload private data for analysis.
- Collaborate securely with colleagues anywhere.

- Use the **400+** cloud- and cost-optimized tools in our Public Apps library.
- Deploy custom tools using our SDK (**Rabix**) & Jupyter notebook (**Data Cruncher**).
- Consult with our **200+** expert support staff.

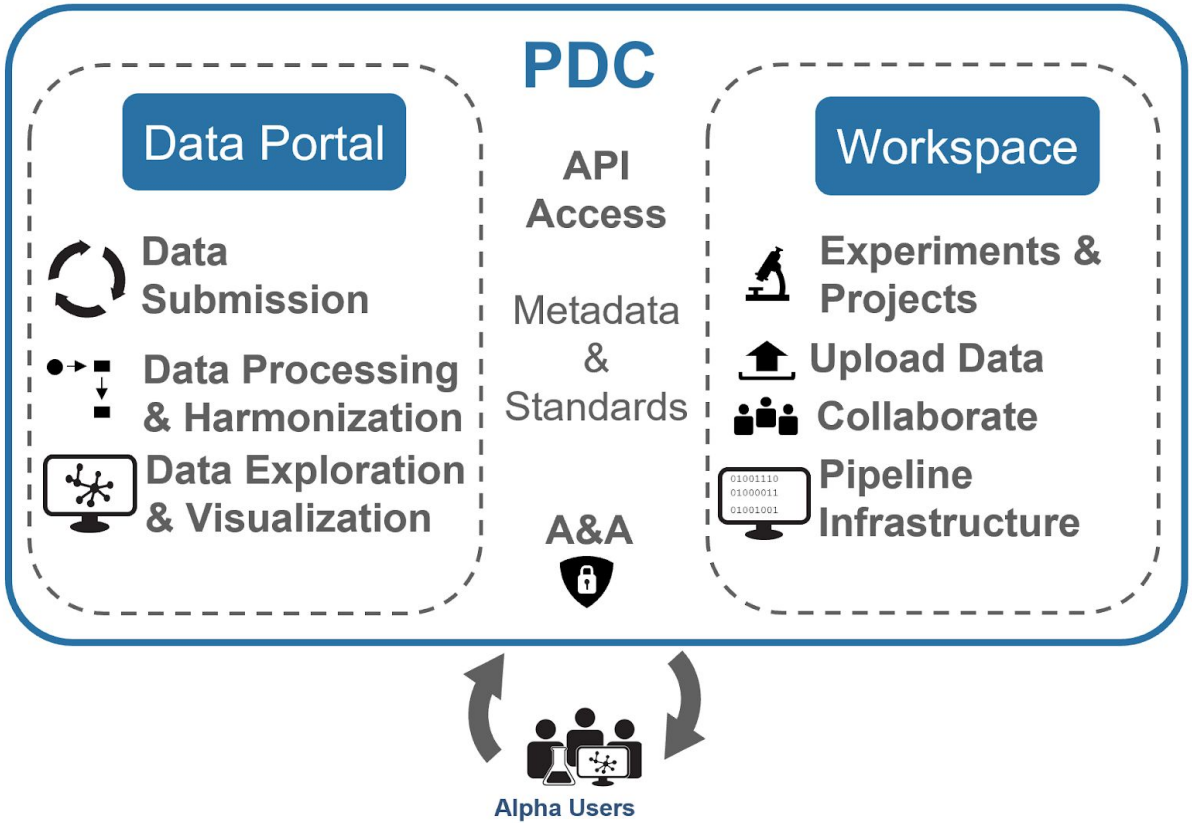




TCGA · The Cancer Imaging Archive · International Cancer Genome Consortium · PCAWG PanCancer Analysis of WHOLE GENOMES · SIMONS FOUNDATION

TARGET · CPTAC · CCLE Cancer Cell Line Encyclopedia

Proteome Data Commons – democratize access to cancer-related proteomic datasets



Courtesy of Izumi HInkson

Enabling multi-omic research on the CGC

- *Many* research questions go beyond genomics data
- Enabling access to the PDC will allow CGC users to **access** and **compute** on data in Proteomics Data Commons in a project on the **CGC**
- Collaboration with Wilson McKerrow, postdoc in David Fenyo's lab
    - Specific use case of analyzing **proteomic, genomic,** and **transcriptomic** data side-by-side in **same project**, ideally for **matching cases** of all 3
    - Useful datasets with all 3 modalities: TCGA, CPTAC-1-

© 2019 Seven Bridges

User flow to compute on PDC files

NATIONAL CANCER INSTITUTE
Proteomic Data Commons

1. User starts on PDC portal to identify cohort of files
2. User downloads **files manifest** of selected cohort
    a. Note - there are multiple manifest options on the PDC, only the “files manifest” will work

CANCER GENOMICS CLOUD
SEVEN BRIDGES

1. User moves to CGC, creates a project
    a. Files ➔ Add files ➔ Import from PDC
2. User prompted to upload the manifest from the PDC
3. PDC files copied into user’s project
    a. FileID ➔ file URL via Fence

© 2019 Seven Bridges

LINE1 retrotransposon project

© 2017 Seven Bridges

© 2017 Seven Bridges