Cloud Computing for Big Data
NCI's CRDC
and
Cloud Resources

Seven Bridges – Cancer Genomics Cloud

Durga Addepalli [CBIIT/NCI]

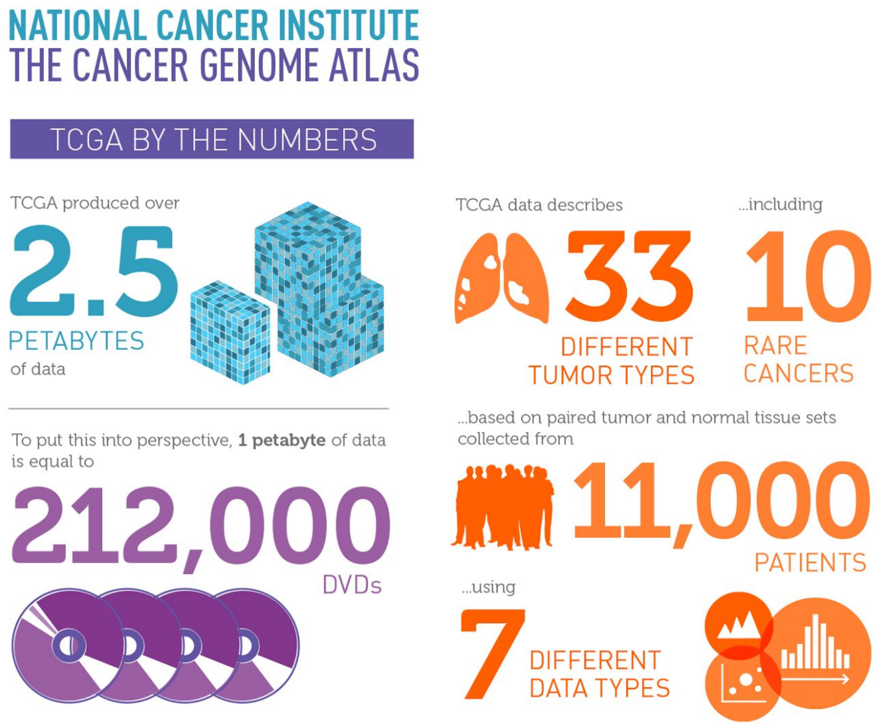
Dave Roberson [SBG-CGC]

BYOB -
10/15/2019

NIH NATIONAL CANCER INSTITUTE

Cancer Genomic Data Challenges

- > 2.5 PB of TCGA data (WXS, RNASeq, WGS)
- Fragmentary repositories of cancer genomic data
  - TCGA, TARGET and CGCI had their own data repositories (DCCs)
  - Sequencing data: BAM files at CGhub while VCF/MAF files at DCC
- Assuming the 2.5 PB TCGA data set
  - Storage and Data Protection would cost approximately $2,000,000 per year
  - Downloading TCGA data at 10 Gb/sec = 23 days
  - Only large institutions could utilize this data
  - These data types continued to grow



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over 2.5 PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to 212,000 DVDs

TCGA data describes 33 DIFFERENT TUMOR TYPES …including 10 RARE CANCERS

…based on paired tumor and normal tissue sets collected from 11,000 PATIENTS

…using 7 DIFFERENT DATA TYPES

F.A.I.R Guiding Principles for Sharing

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

T
s
c
t
e
s
f
f
i

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

PERSPECTIVE

Check for updates

Downloaded from mbio.asm.org on June 19, 2018 - Published by mbio.asm.org

Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research

Patrick D. Schloss[a]

[a]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA

ABSTRACT The "reproducibility crisis" in science affects microbiology as much as any other area of inquiry, and microbiologists have long struggled to make their research reproducible. We need to respect that ensuring that our methods and results are sufficiently transparent is difficult. This difficulty is compounded in interdisciplinary fields such as microbiome research. There are many reasons why a researcher is unable to reproduce a previous result, and even if a result is reproducible, it may not be correct. Furthermore, failures to reproduce previous results have much to teach us about the scientific process and microbial life itself. This Perspective delineates a framework for identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability of microbiome research. Instead of seeing signs of a crisis in others' work, we need to appreciate the technical and social difficulties that limit reproducibility in the work of others as well as our own.

Sharing Analysis - Tools and Workflows?

Container/Workflow Technologies

COMMON WORKFLOW LANGUAGE

docker

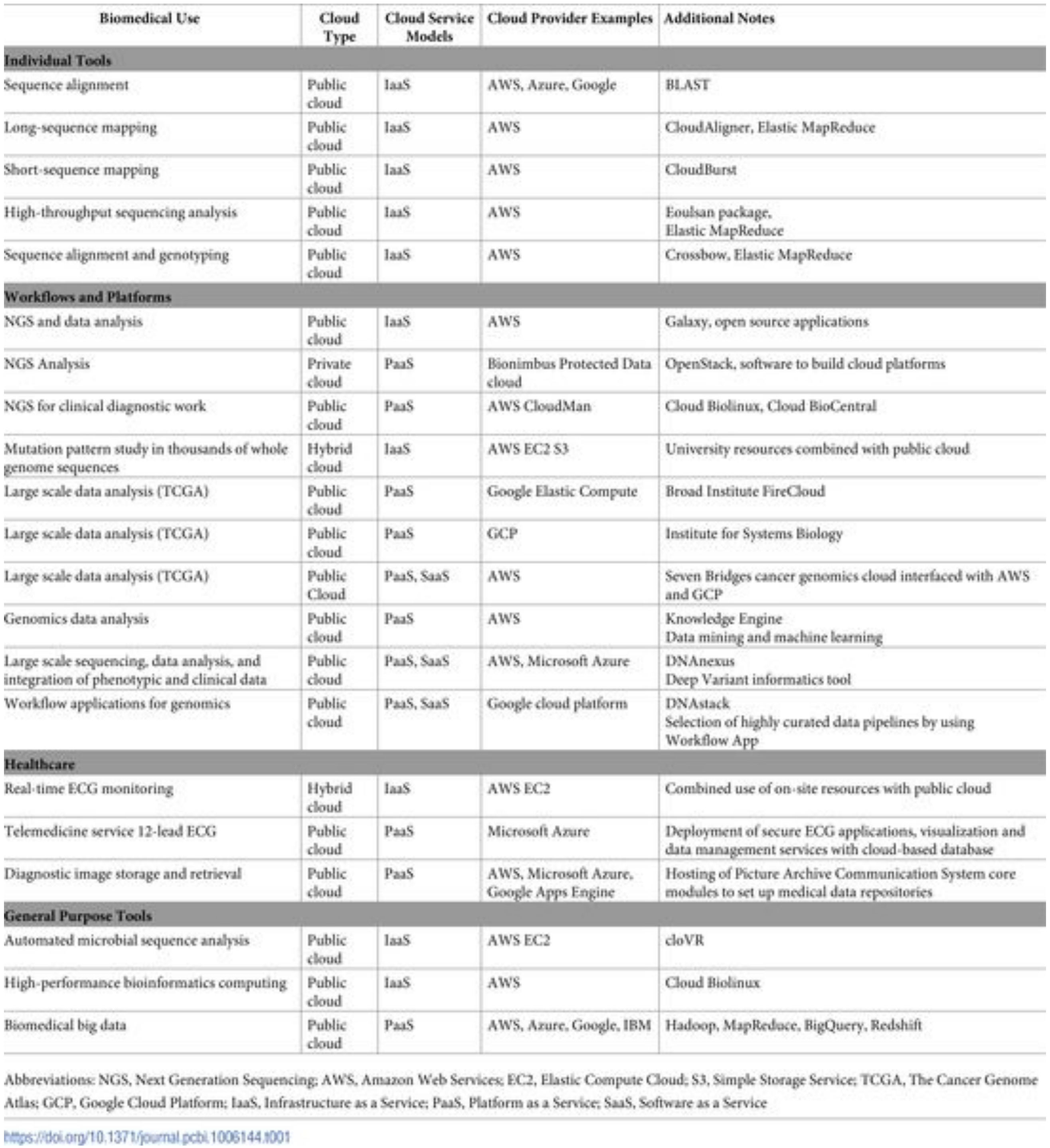
{wdl}

nextflow

NIH NATIONAL CANCER INSTITUTE

Examples of cloud types, service models, workflows, and platforms for biomedical applications.

Navale V, Bourne PE (2018) Cloud computing applications for biomedical science: A perspective. PLOS Computational Biology 14(6): e1006144. https://doi.org/10.1371/journal.pcbi.1006144
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006144



Biomedical Use	Cloud Type	Cloud Service Models	Cloud Provider Examples	Additional Notes
Individual Tools				
Sequence alignment	Public cloud	IaaS	AWS, Azure, Google	BLAST
Long-sequence mapping	Public cloud	IaaS	AWS	CloudAligner, Elastic MapReduce
Short-sequence mapping	Public cloud	IaaS	AWS	CloudBurst
High-throughput sequencing analysis	Public cloud	IaaS	AWS	Eoulsan package, Elastic MapReduce
Sequence alignment and genotyping	Public cloud	IaaS	AWS	Crossbow, Elastic MapReduce
Workflows and Platforms				
NGS and data analysis	Public cloud	IaaS	AWS	Galaxy, open source applications
NGS Analysis	Private cloud	PaaS	Bionimbus Protected Data cloud	OpenStack, software to build cloud platforms
NGS for clinical diagnostic work	Public cloud	PaaS	AWS CloudMan	Cloud Biolinux, Cloud BioCentral
Mutation pattern study in thousands of whole genome sequences	Hybrid cloud	IaaS	AWS EC2 S3	University resources combined with public cloud
Large scale data analysis (TCGA)	Public cloud	PaaS	Google Elastic Compute	Broad Institute FireCloud
Large scale data analysis (TCGA)	Public cloud	PaaS	GCP	Institute for Systems Biology
Large scale data analysis (TCGA)	Public Cloud	PaaS, SaaS	AWS	Seven Bridges cancer genomics cloud interfaced with AWS and GCP
Genomics data analysis	Public cloud	PaaS	AWS	Knowledge Engine Data mining and machine learning
Large scale sequencing, data analysis, and integration of phenotypic and clinical data	Public cloud	PaaS, SaaS	AWS, Microsoft Azure	DNAnexus Deep Variant informatics tool
Workflow applications for genomics	Public cloud	PaaS, SaaS	Google cloud platform	DNAstack Selection of highly curated data pipelines by using Workflow App
Healthcare				
Real-time ECG monitoring	Hybrid cloud	IaaS	AWS EC2	Combined use of on-site resources with public cloud
Telemedicine service 12-lead ECG	Public cloud	PaaS	Microsoft Azure	Deployment of secure ECG applications, visualization and data management services with cloud-based database
Diagnostic image storage and retrieval	Public cloud	PaaS	AWS, Microsoft Azure, Google Apps Engine	Hosting of Picture Archive Communication System core modules to set up medical data repositories
General Purpose Tools				
Automated microbial sequence analysis	Public cloud	IaaS	AWS EC2	cloVR
High-performance bioinformatics computing	Public cloud	IaaS	AWS	Cloud Biolinux
Biomedical big data	Public cloud	PaaS	AWS, Azure, Google, IBM	Hadoop, MapReduce, BigQuery, Redshift

Abbreviations: NGS, Next Generation Sequencing; AWS, Amazon Web Services; EC2, Elastic Compute Cloud; S3, Simple Storage Service; TCGA, The Cancer Genome Atlas; GCP, Google Cloud Platform; IaaS, Infrastructure as a Service; PaaS, Platform as a Service; SaaS, Software as a Service

https://doi.org/10.1371/journal.pcbi.1006144.t001

NCI Cancer Research Data Commons (CRDC)

What is data commons?

Data commons co-locate data, storage and computing infrastructure with commonly used tools for analyzing and sharing data to create an interoperable resource for the research community.

Goals of the NCI CRDC:

- Enable the cancer research community to share diverse data types across programs and institutions
- Provide secure access to data, regardless of where it is stored
- Provide mechanisms for innovative tool discovery, visualization, analysis using elastic compute
- Help NCI Data Coordinating Centers sustain and share their data publicly

NCI CRDC Components

Data Commons Framework

Data Nodes

Cloud Resources

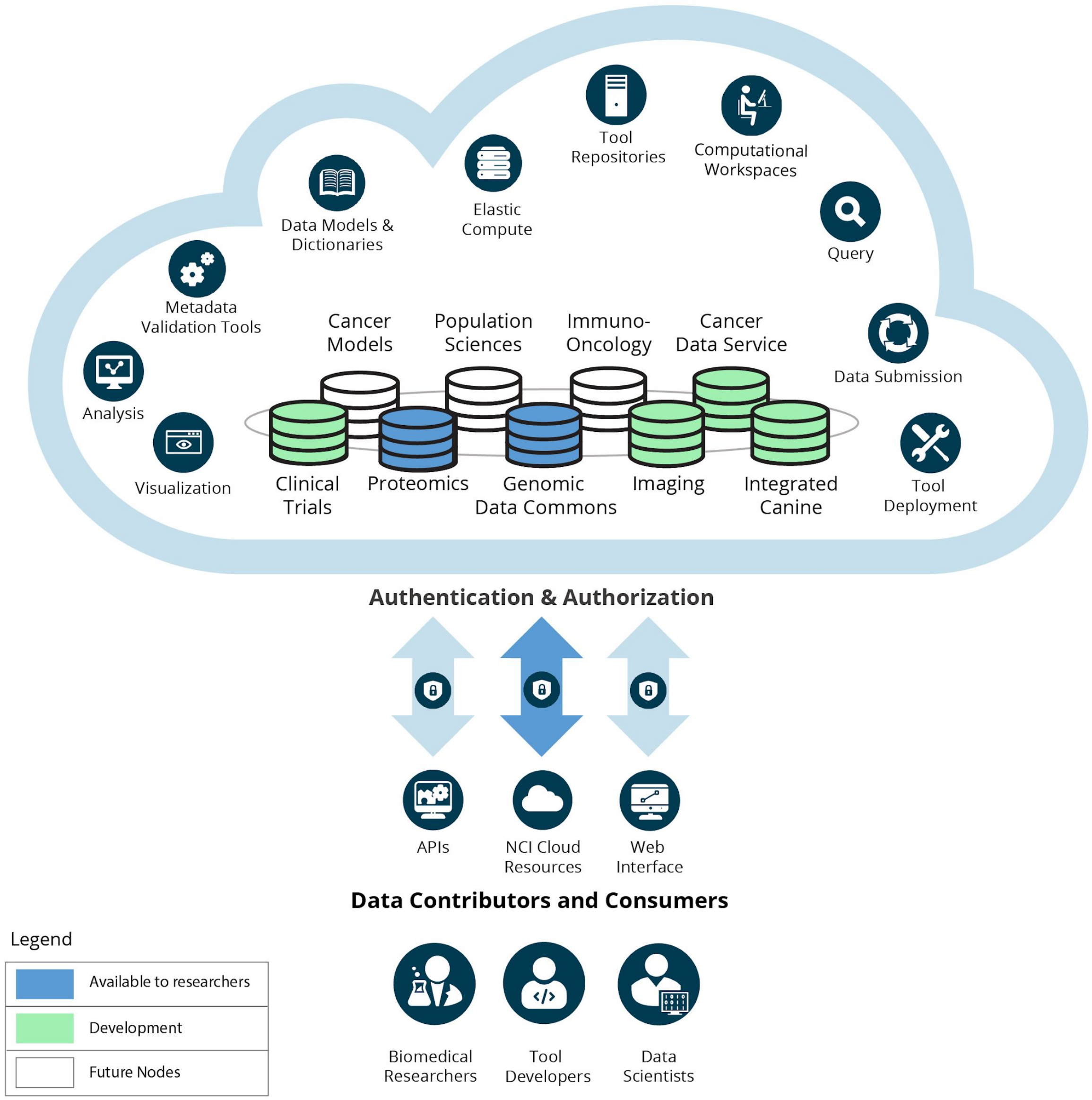
Semantics Support

Cancer Data Aggregator

NCI CRDC

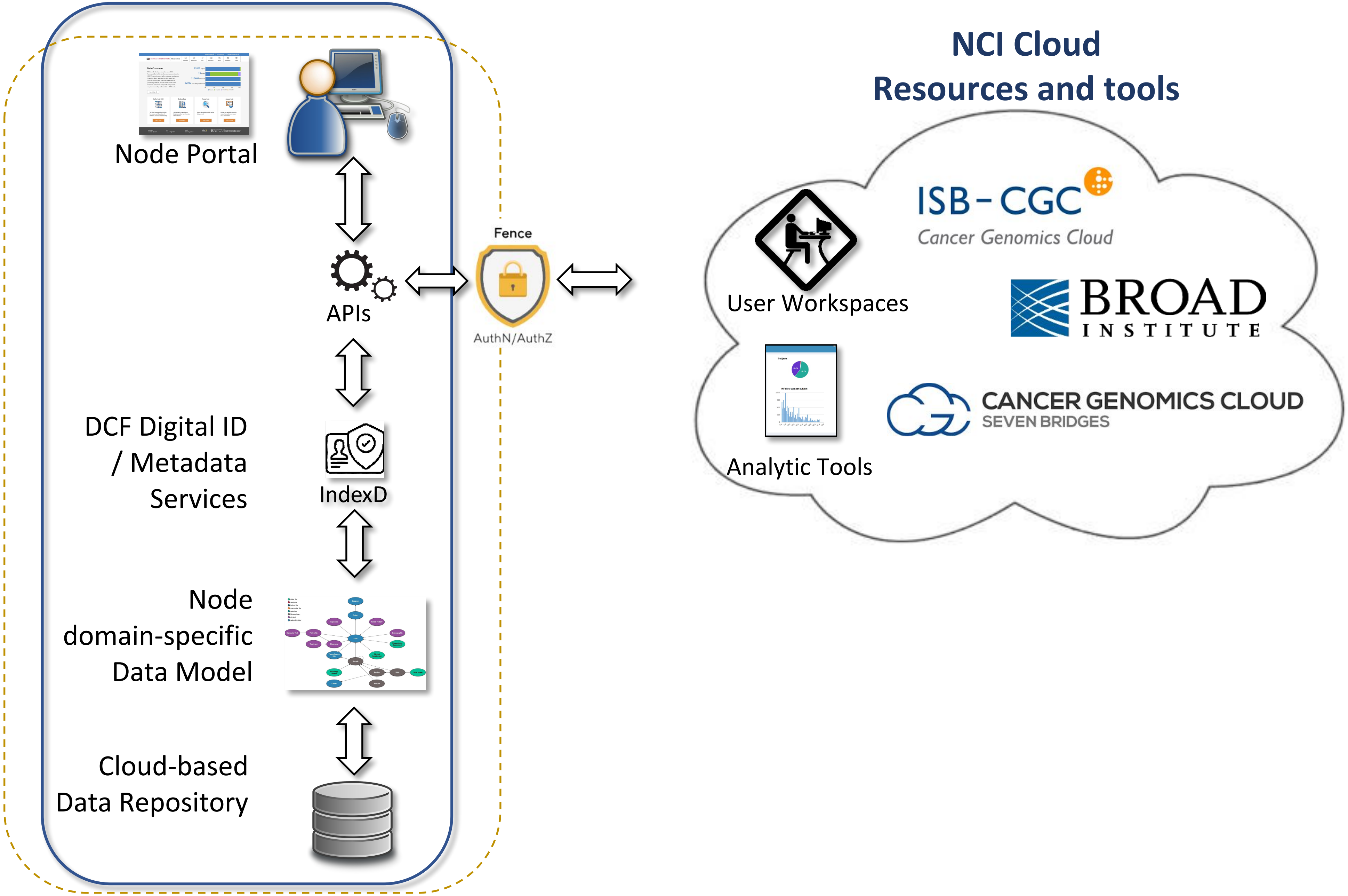
- Data are stored in domain-specific repositories, called Data Nodes (e.g., genomic, proteomic, imaging, etc.).
- Data access is controlled through a common Authentication and Authorization mechanism that secures the data.
- Researchers can bring their own data and tools to the cloud, and combine with the data in the CRDC for integrative analysis.

NCI Cancer Research Data Commons (CRDC)



CRDC Node

NCI Cloud Resources and tools



Node Portal

APIs

Fence

AuthN/AuthZ

DCF Digital ID / Metadata Services

IndexD

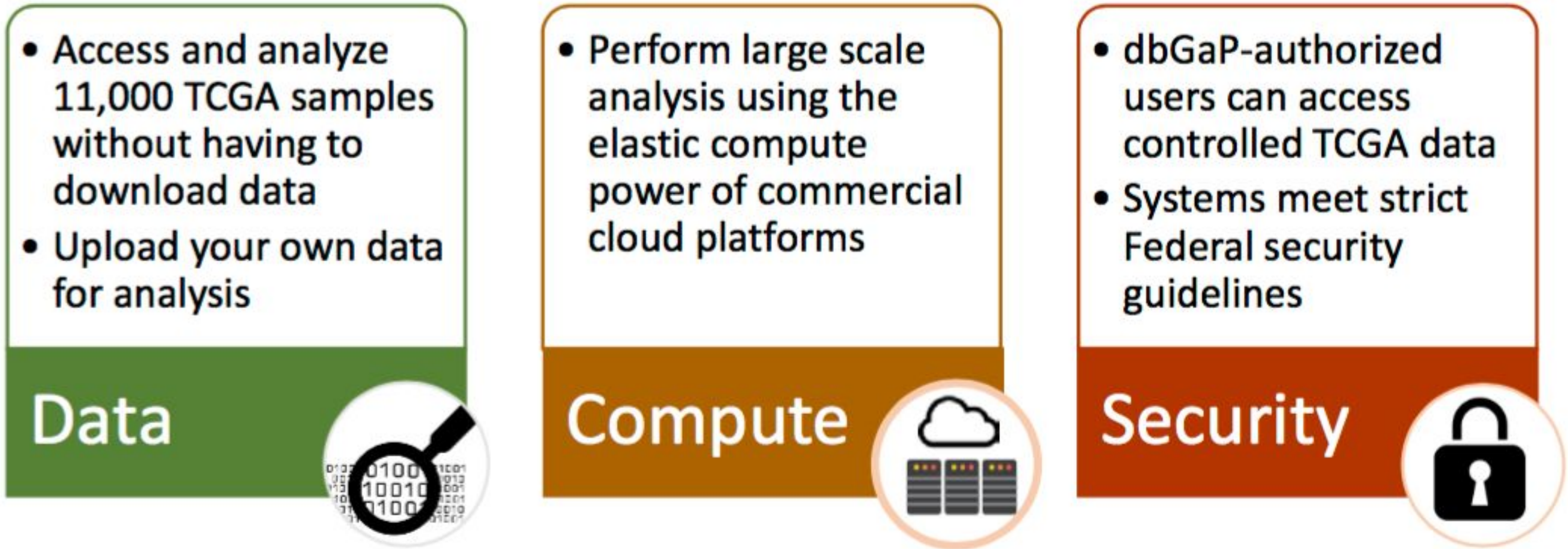
Node domain-specific Data Model

Cloud-based Data Repository

User Workspaces

Analytic Tools

NCI Cloud Resources



- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Data

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute

- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Security

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

http://cgc.systemsbiology.net/

CANCER GENOMICS CLOUD
SEVEN BRIDGES

http://www.cancergenomicscloud.org

FireCloud POWERED BY Terra

https://firecloud.terra.bio/#

The Cloud Resources provide:

- Access to large cancer data sets without need to download
- Access to popular analysis tools and pipelines
- Ability for researchers to bring their own data to the Cloud Resources
- Ability for researchers to bring their own tools and pipelines to the data
- Workspaces, for researchers to save and share their data and results of analyses

Three NCI Cloud Resources

Broad Institute

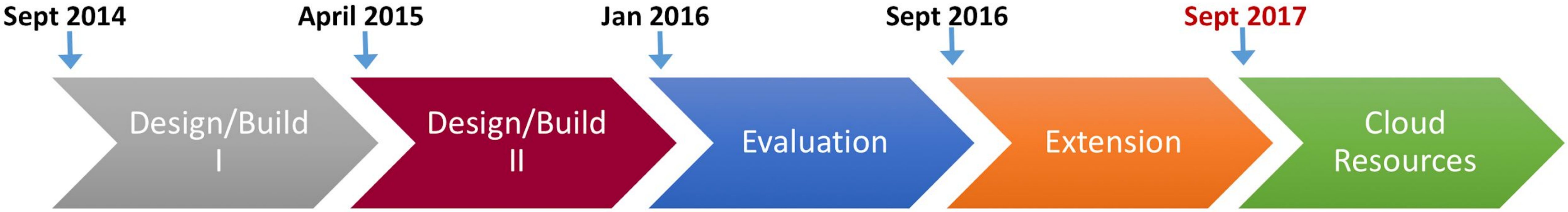
- PI: Anthony Philippakis
- Google Cloud
- Firehose in the cloud including Broad best practices workflows
- http://firecloud.org

Institute for Systems Biology

- PI: Bill Longabaugh
- Google Cloud
- Leverage Google infrastructure; Novel query and visualization
- https://isb-cgc.appspot.com

Seven Bridges Genomics

- PI: Brandi Davis-Dusenbery
- Amazon Web Services
- Interactive data exploration; > 30 public pipelines
- http://www.cancergenomicscloud.org



Sept 2014 – Design/Build I

April 2015 – Design/Build II

Jan 2016 – Evaluation

Sept 2016 – Extension

Sept 2017 – Cloud Resources

Data Sets available to the Cloud Resources

Each new data connection brings more analysis possibilities

- TCGA and TARGET
- CTSP (Clinical Trials Sequence Project)
- FM (Foundation Medicine)
- NCICCR - Diffuse large B cell lymphomas (DLBCL) study
- VAREPOP (VA APOLLO Project - Research for Precision Oncology (RePOP))
- CCLE (Cancer Cell Line Encyclopedia)
- CPTAC2 and CPTAC3 - genomic and proteomic data sets
- MMRF - genomic and clinical data on multiple myeloma patients
- GECCO (Genetics and Epidemiology of Colorectal Cancer Consortium)
- Canine Data (clinical trials, genomic, imaging, immunology, ect)
- IDC: Imaging Data Commons (TCIA-The Cancer Imaging Archive, HTAN- Human Tumor Atlas. Network, TBD ….)
- APOLLO: The ***A***pplied ***P***roteogenomics ***O***rganizationa***L*** ***L***earning and ***O***utcomes
- ICPC: The International Cancer Proteogenome Consortium
  ………….. (many more to come)

Links

- NCI CRDC general info
  - https://datascience.cancer.gov/data-commons
- Data Nodes
  - GDC: http://gdc.cancer.gov
  - PDC: http://pdc.esacinc.com/pdc
- Data Commons Framework
  - http://dcf.gen3.org

- Cloud Resources
  - Broad FireCloud: http://firecloud.org
  - Seven Bridges CGC: http://www.cancergenomicscloud.org
  - Institute for Systems Biology CGC: http://cgc.systemsbiology.net

- Contact:
  - CRDC & CRs- Durga Addepalli kanakadurga.addepalli@nih.gov
  - SBG CGC - Dave Roberson david.roberson@sevenbridges.com