The Cancer Data Ecosystem
The Beau Biden Cancer Moonshot℠

Overarching goals – Jan, 2016

• Accelerate progress in cancer, including prevention & screening
  • From cutting edge basic research to wider uptake of standard of care

• Encourage greater cooperation and collaboration
  • Within and between academia, government, and private sector

• Enhance data sharing

Blue Ribbon Panel – October, 2016

• Network for Direct Patient Engagement
• Cancer Immunotherapy Translational Science Network
• Therapeutic Target Identification to Overcome Drug Resistance
• A National Cancer Data Ecosystem for Sharing and Analysis
• Fusion Oncoproteins in Childhood Cancers
• Symptom Management Research
• Prevention and Early Detection – Implementation of Evidence-based Approaches
• Retrospective Analysis of Biospecimens from Patients Treated with Standard of Care
• Generation of 3D Human Tumor Atlas
• Development of New Enabling Cancer Technologies
• Full report:  www.cancer.gov/brp
National Cancer Data Ecosystem Recommendations

Overall goal: “Enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer.”

Recommendations

• Build a National Cancer Data Ecosystem
  • Enhanced cloud-computing platforms.
  • Services that link disparate information, including clinical, image, and molecular data.
  • Essential underlying data science infrastructure, methods, and portals for the Cancer Data Ecosystem.
  • Establish sustainable data governance to ensure long-term health of the Ecosystem.
  • Develop standards and tools so that data are interoperable.
Enhanced Data Sharing Working Group Recommendation:

**The Cancer Data Ecosystem**
The Cancer Research Data Commons (CRDC)
NCI Scope: “Create a data science infrastructure necessary to connect repositories, analytical tools, and knowledge bases”

Data commons co-locate data, storage and computing infrastructure with commonly used services, tools & apps 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 easy access to data, regardless of where they are stored.

• Provide mechanisms for innovative tool discovery, access, and usage, e.g., ITCR tools.

• Help Data Coordinating Centers share their data publicly.
NCI Cancer Research Data Commons
Data Sources / Contributors (Examples)

- The Cancer Genome Atlas (TCGA)
- Clinical Proteomic Tumor Analysis Consortium (CPTAC)
- The Cancer Imaging Archive (TCIA)
- NCI Individual Labs / Grants / Contracts / Cancer Centers (GENIE)
- Therapeutically Applicable Research to Generate Effective Treatments (TARGET)

Collaborative Programs: APOLLO (Applied Proteogenomic Organizational Learning and Outcomes), ICPC (International Cancer Proteogenome Consortium)

3rd Party Programs: Foundation Medicine, Multiple Myeloma Research Foundation
Data Commons Framework – What Is It?

**Modular Components**

<table>
<thead>
<tr>
<th>Icon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>🔒</td>
<td>Secure user authentication and authorization</td>
</tr>
<tr>
<td>🔍</td>
<td>Metadata validation and tools</td>
</tr>
<tr>
<td>📖</td>
<td>Domain-specific, extensible data models and dictionaries</td>
</tr>
<tr>
<td>🗂</td>
<td>API and container environment for tools and pipelines</td>
</tr>
<tr>
<td>📝</td>
<td>Access to computational workspaces for storing data, tools, and results</td>
</tr>
</tbody>
</table>

Reusable, expandable framework for a Data Commons

Core principles and structures for a Data Commons

Set of modular components that can be leveraged across Data Commons
NCI Cancer Research Data Commons

Data Commons Framework

Authentication & Authorization

APIs

Web Interface

Data Submission

Tool Deployment

Authentication & Authorization

Data Contributors and Consumers

NCI Cancer Research Data Commons

Animal Models

Clinical

Genomics

Proteomics

Imaging

Cancer Biomarkers

Visualization

Elastic Compute

Computational Workspaces

Data Models & Dictionaries

NCI Cancer Research Data Commons

CPTAC
Clinical Proteomics Tumor Analysis Consortium*

TCIA
The Cancer Imaging Archive*

Tool Repositories

SBG CGC
Broad FireCloud
ISB CGC

Metadata Validation & Tools

Data Contributors and Consumers

Biomedical Researchers

Tool Developers

Computer Scientists

Clinicians

Patients
The NCI Genomic Data Commons

Provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.
The NCI Genomic Data Commons

• Support the receipt, quality control, integration, storage, and redistribution of standardized genomic data sets derived from cancer research studies
  • Available data
    • NCI Funded cancer genomics datasets
    • User submissions
  • Data searching and retrieval/downloading
  • Harmonization of raw sequence (alignment and variant calling) of all GDC data
    • Application of state-of-the-art methods of generating derived data
• Developed, supported, and hosted by U. Chicago
Genomic Data Commons (GDC):
A unified data repository for the research community developed, supported, and hosted by U. Chicago

Genomic Data Commons:
Harmonization, Visualization, & Download

Data Submission & Harmonization

Web Interface

APIs

Researchers

https://gdc.cancer.gov
GDC: Data Submission & Harmonization

Data Submission

Data Harmonization

https://gdc.cancer.gov/
GDC: Data Retrieval

Data Portal

Data Transfer Tool

GDC Website

Visualization Tools

GDC API

```
"data": {
"hits": [
{"project_id": "TCGA-SKCM", "primary_site": "Skin"},
{"project_id": "TCGA-PCPG", "primary_site": "Nervous System"},
{"project_id": "TCGA-LAML", "primary_site": "Blood"},
{"project_id": "TCGA-UVM", "primary_site": "Eye"}
]
}
```

API URL: https://gdc-api.nci.nih.gov/projects?fields=project_id,primary_site&pretty=true

GDC Legacy Archive

GDC: Data Retrieval

API URL: https://gdc-api.nci.nih.gov/projects?fields=project_id,primary_site&pretty=true

Visualization Tools

GDC: Data Retrieval

GDC Website

Visualization Tools
The NCI Cloud Resources

Understanding how to meet the research community’s need to analyze large-scale cancer genomic and clinical data
GDC and the NCI Cloud Resources

Genomic Data Commons:
- Harmonization,
- Visualization,
- Download

Cloud Resources:
- Compute, Pipelines,
- Workspaces

Authentication & Authorization thru eRA Commons & dbGaP

Data Submission & Harmonization

APIs

Web Interface

Researchers

Compute, Pipelines, Workspaces

SBG CGC

Broad FireCloud

ISB CGC
NCI Cloud Resources

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

Democratize access to cancer datasets, and to provide cost-effective computational capacity to the cancer research community.
Three NCI Cloud Resources

**Broad Institute**
- PI: Anthony Philippakis
- Google Cloud
- Firehose in the cloud including Broad best practices workflows
  - [http://firecloud.org](http://firecloud.org)

**Institute for Systems Biology**
- PI: Ilya Shmulevich
- Google Cloud
- Leverage Google infrastructure; Novel query and visualization
  - [http://cgc.systemsbiology.net/](http://cgc.systemsbiology.net/)

**Seven Bridges Genomics**
- PI: Brandi Davis-Dusenbery
- Amazon Web Services
- Interactive data exploration; > 30 public pipelines
  - [http://www.cancergenomicscloud.org](http://www.cancergenomicscloud.org)
Broad Institute Cloud Resource

• Targeted at users performing analyses at scale.
• Modeled after their Firehose analysis infrastructure developed for the TCGA program.
• Users can upload their own data and tools and/or run the Broad’s best practice tools and pipelines on pre-loaded data.

http://firecloud.org
The Data Library

The primary tool for discovering datasets at Broad and beyond

Broad’s Genomics Platform has been delivering all WGS projects into FireCloud for a year.

Recently begun cataloguing all data into the Data Library for discovery and access.

Users can search the datasets and filter datasets by the data use restrictions.
The $5 Genome: Pipeline Optimizations

- Broad is optimizing production pipelines with a commitment to openness, transparency, and continued improvements in cost and performance
- Example: Germline GATK best practices
  - $45/sample* in 2016
  - $13.50/sample in 2017
  - $5/sample in 2018
- Pipelines will be available to run in FireCloud and will also be in Dockstore.

Optimized somatic best practices coming soon!

* Cloud compute costs from Google Cloud Platform
FireCloud is part of the Data Biosphere

FireCloud will evolve into a citizen of an interoperable world through principles outlined across the Data Biosphere.

The Biosphere is a collaboration among institutions working on data platforms that will serve several large-scale, high-profile biomedical research projects.

Principles

- Open
- Standards Based
- Modular
- Community Driven

Initial collaborators in the Data Biosphere are building data platforms for the NCI Data Commons, NIH Data Commons, All of Us Research Program, Human Cell Atlas, Gabriella Miller Kids First, and others.
First integration with Data Biosphere: Dockstore

Dockstore workflows on Dockstore.org can now be launched with FireCloud!
ISB Cancer Genomics Cloud (ISB-CGC)

- Closely tied with Google Cloud Platform tools including BigQuery, App Engine, Cloud Datalab, Google Genomics, and Compute Engine

http://cgc.systemsbiology.net/
There are three primary ways of working on ISB-CGC.

**ISB-CGC Web App**

**BigQuery**

**Pipelines**

What you choose depends on the question and what you're comfortable with.
Method 2: Working with BigQuery

- bigrquery and bigQueryR
- googleAuthR
- Bioconductor: Pre-built VM images

Cloud Datalab

- IP[y]: IPython and Jupyter
- Cloud notebooks and workspaces.

Google BigQuery plays well with others.
it's great for answering questions

Q: How many samples have a mutation in PARP1?

Use standard SQL to answer it:

```
SELECT
    project_short_name,
    COUNT(DISTINCT(sample_barcode_tumor)) AS n
FROM
`isb-cgc.TCGA_hg38_data_v0.Somatic_Mutation_DR10`
WHERE
    Hugo_Symbol = 'PARP1'
GROUP BY
    project_short_name
ORDER BY
    n DESC
```

Easy to join tables on any shared variable.

Lots of built in functions for math, string processing, etc.

Can process massive amounts of data in parallel.
Query Of The Month Club

Spearman correlations using RNA-seq data and pathway definitions

ISB-CGC Query of the Month, Feb 2018

This plot shows gene-gene correlations for a set of genes given by the selected pathway. The correlations can be filtered using the correlation threshold slider. BioCircos links can be moused-over to display the gene pair and correlation value. Also, it’s possible to zoom in on portions of the circos plot by double-clicking. Try searching the list of pathways by selecting the pathway drop-down, hitting delete and typing a search term.

Pathway
CELL CYCLE CHECKPOINTS (291 genes)

Cohort
TCGA-LUSC

random number of genes

1 2 3 4 5 6 7 8 9 10

correlation threshold

Submit after selecting pathway and cohort

Packages used: BioCircos, bigquery
Data used: Reactome pathways, TCGA hg19 RNAseq UNC RSEM
ISB-CGC a key resource for TCGA #PanCancerAtlas

- Germline, Fusion, and Immune Response papers used ISB-CGC to access and compute on TCGA sequence data on the Google Cloud Platform
- Immune, MYC, and DDR papers used BigQuery and ISB-CGC data tables
- #PanCancerAtlas open-access tables now available in BigQuery (referenced from GDC page)
- The availability of PanCaner Atlas data in BigQuery enables easy integration with other public datasets through BigQuery
The Seven Bridges Cancer Genomics Cloud (CGC)

- A user-friendly, web-based portal for collaborative analysis of petabytes of multi-omic data alongside private data
- Built upon the SBG commercial cloud-based genomics platform
- For cancer genomics research and beyond

易用的数据管理
可扩展的计算
安全的合作
优化的生物信息学算法
灵活且完全可重复的方法
可扩展且开发者友好的平台

http://www.cancergenomicscloud.org
Available Resources

- 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 360\(^+\) cloud- and cost-optimized tools in the Public Apps library.
- Deploy custom tools using SDK (Rabix) & Jupyter notebook (Data Cruncher).
- Consult with 200\(^+\) expert support staff.
Usage by the Research Community

3,100+ users from 60+ countries have used the CGC to run 347,000+ computational tasks representing 465+ years of total compute time to:

- Detect aberrant splice junctions and splicing profiles across patient populations
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across patient populations
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across patient populations
- Detect novel gene fusions from RNA-Seq data
- Identify cis-regulatory region variants across patient populations
- ...and much more
Scalable, Cost-Effective Research

Case Study #1: TCGA Immune Response Working Group
- Collaborative analysis with members of the Immune Response Working Group of The Cancer Genome Atlas (TCGA) Research Network
- Outcome: cost-optimized (<$0.30/sample), high-throughput HLA typing across ~9,000 TCGA RNA-Seq (fastq) files

Case Study #2: PanCancer Analysis of Whole Genomes (PCAWG) Study from International Cancer Genome Consortium (ICGC)
- High-throughput, harmonized analysis by Seven Bridges of all tumor and matched genomes in the dataset (~1,350)
- Outcome: rapid generation of ~65,000 output files (including ~5,000 VCFs) totaling 725 TB

Case Study #3: Independent Analysis on 45,000 Bacterial Genomes
- High-throughput analysis of 45,000 bacterial genomes accessed from SRA via API and analyzed using a custom workflow
- Outcome: analysis completed in ~1 week by a novice CGC user with no substantive assistance from the CGC team