1. National Cancer Data Ecosystem

2. NCI Cancer Research Data Commons

3. Data Linkages
   - Cancer Data Aggregator
   - Encrypted Unique Patient Identifier

4. Collaboration/coordination
   - Partnerships
   - Office of Data Sharing
1. National Cancer Data Ecosystem

2. NCI Cancer Research Data Commons

3. Data Linkages
   - Cancer Data Aggregator
   - Encrypted Unique Patient Identifier

4. Collaboration/coordination
   - Partnerships
   - Office of Data Sharing
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
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.
  • Essential underlying data science infrastructure and portals for the Cancer Data Ecosystem.
  • Services that link disparate information, including clinical, image, and molecular data.
  • Develop standards and tools so that data are interoperable.
  • Address sustainability and data governance to ensure long-term health of the Ecosystem.

• The National Cancer Data Ecosystem is broader than NCI
  • An NCI Cancer Research Data Commons is envisioned as part of the National Cancer Data Ecosystem
Enhanced Data Sharing Working Group Recommendation:

The Cancer Data Ecosystem
Agenda

1. National Cancer Data Ecosystem

2. NCI Cancer Research Data Commons

3. Data Linkages
   - Cancer Data Aggregator
   - Encrypted Unique Patient Identifier

4. Collaboration/coordination
   - Partnerships
   - Office of Data Sharing
NCI Cancer Research Data Commons (CRDC) - Concept

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 it is stored.

• Provide mechanisms for innovative tool discovery, access, usage.

• Help NCI Data Coordinating Centers sustain and share their data publicly.

• Develop a set of reusable components - a framework - for the community to use to build interoperable data commons.
CRDC Data Sources / Contributors (Examples)

- **The Cancer Genome Atlas (TCGA)**
- **Therapeutically Applicable Research to Generate Effective Treatments (TARGET)**
- **3rd Party Programs: Foundation Medicine, Multiple Myeloma Research Foundation**
- **Clinical Proteomic Tumor Analysis Consortium (CPTAC)**
- **The Cancer Imaging Archive (TCIA)**
- **NCI Individual Labs / Grants / Contracts / Cancer Centers (GENIE)**
- **Collaborative Programs: APOLLO (Applied Proteogenomic Organizational Learning and Outcomes), ICPC (International Cancer Proteogenome Consortium)**
Data Commons Framework – What Is It?

### Modular Components

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Lock Icon]</td>
<td>Secure user authentication and authorization</td>
</tr>
<tr>
<td>![Gear Icon]</td>
<td>Metadata validation and tools</td>
</tr>
<tr>
<td>![Book Icon]</td>
<td>Domain-specific, extensible data models and dictionaries</td>
</tr>
<tr>
<td>![Container Icon]</td>
<td>API and container environment for tools and pipelines</td>
</tr>
<tr>
<td>![Person Icon]</td>
<td>Access to computational workspaces for storing data, tools, and results</td>
</tr>
</tbody>
</table>

NCI is developing the Framework and will use it to stand up several example Data Commons the community can leverage or use as a model to build their own commons.
1. NCI GDC / NCRDC (UChicago)
2. NIH All of Us (Broad/Verily)
3. CZI HCA Data Platform (UCSC/Broad)

For more information, see: Josh Denny, David Glazer, Robert L. Grossman, Benedict Paten & Anthony Philippakis, A Data Biosphere for Biomedical Research, [https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d](https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d). Also available at: [https://goo.gl/9CySeo](https://goo.gl/9CySeo)
Narrow Middle Architecture (End-to-End Design)

1. AuthN / AuthZ
2. Metadata validation
3. Extensible data model
4. APIs for containers, workflows & tools
5. Workspaces

Building on Foundation of the NCI Genomic Data Commons

From Grossman et al. 
N Engl J Med 2016; 375:1109-1112
September 22, 2016 
DOI: 10.1056/NEJMp1607591

https://gdc.cancer.gov/
NCI Cloud Resources

Cloud Resources provide:
- Access to large genomic data sets without need to download
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

#NCICloud
GDC and Cloud Resources are available now; Framework, As Is Genomics, PDC, IDC are in development; all else is notional.
Agenda

1. National Cancer Data Ecosystem
2. NCI Cancer Research Data Commons
3. Data Linkages
   - Cancer Data Aggregator
   - Encrypted Unique Patient Identifier
4. Collaboration/coordination
   - Partnerships
   - Office of Data Sharing
Goal: Provide a reusable informatics service to connect disparate data in support of integrative cancer research.
Development of an Encrypted Unique Patient Identifier

• Pressing need to connect patient-level data across multiple data sources, data types and research studies—over time.

• Challenges include:
  • Protecting patient confidentiality
  • Consistency of identifying data (personally identifiable information, PII) available across diverse sources
  • Accuracy of linkage with varying PII
  • Scalability

• Encrypted hashed token
  • Allows linkage of diverse data.
  • Permits data sharing across multiple sources without release of PII.
Agenda

1. National Cancer Data Ecosystem
2. NCI Cancer Research Data Commons
3. Data Linkages
   - Cancer Data Aggregator
   - Encrypted Unique Patient Identifier
4. Collaboration/coordination
   - Partnerships
   - Office of Data Sharing
Creating Partnerships

• Administrative supplements for Cancer Centers in GENIE and GA4GH coordination.
• Coordination with and support of Moonshot Programs
  • Assistance for U24 programs, e.g., Human Tumor Atlas & Immuno-oncology Data Coordinating Centers
• Work across related initiatives/programs
  • NCI, other NIH Institutes, NIH Data Commons Pilot Phase Consortium, All of Us, Chan Zuckerberg Initiative, GA4GH
• Establishing NCI Office of Data Sharing as a resource to NCI staff and extramural investigators.
• Workshops and RFIs to gather community input, feedback, and participation
  • Semantics infrastructure workshop
• Establish CRDC governance process, including Scientific and Technical Advisory Board and Steering Committee.
Office of Data Sharing Activities

- **Coordinates** the interpretation and implementation of data sharing policies across NCI
- Provides **workflow management** and coordination of NCI data/metadata submissions and access processes relative to NIH databases, including dbGaP
- **Advocates** for the proper balance of open access, open source, broad data sharing policies
- **Outreach and education** on NCI data sharing policies and processes; central clearing house for knowledge management
- Develops and monitors **metrics** relevant for understanding influence, uptake, and compliance regarding NCI data/metadata usage
- Coordinates with and provides leadership as appropriate to other key organizations within NIH and the research community
Office of Data Sharing Tentative Schedule

Jan – May 2018

Begin process of standing up the NCI Office of Data Sharing within CBIIT (approve organizational structure and begin to hire key personnel)

June – Sept 2018

Work with NCI Divisions/Offices/Centers to develop a plan for centralizing current NCI Data Sharing activities

Oct – Dec 2018

Implement centralization plan for NCI Data Sharing activities moving to the ODS

January 2019 –

Work within NCI and the NIH, as well as the external research community, to best implement NIH GDS and other Data Sharing policies as relevant to cancer data and metadata
Cancer Research Data Commons Project Teams

**CRDC Framework Principal Investigators**
- Robert Grossman - University of Chicago
- Anthony Philippakis - Broad Institute
- Ilya Shmulevich – Institute for Systems Biology
- Brandi Davis-Dusenbery - Seven Bridges

**CBIIT Data Commons Team**
- Tanja Davidsen
- Ian Fore
- Izumi Hinkson
- Betsy Hsu
- Steve Jett
- Tony Kerlavage
- Juli Klemm
- David Patton

**Surveillance Research Program**
- Lynne Penberthy
- Paul Fearn

**Center for Cancer Genomics**
- Lou Staudt
- JC Zenklusen
- Daniela Gerhard
- Zhining Wang
- Liming Yang
- Martin Ferguson

**Center for Strategic Scientific Initiatives**
- Chris Kinsinger
- Henry Rodriguez

**Cancer Imaging Program**
- Paula Jacobs
- John Freymann
- Justin Kirby

**Leidos Biomedical Data Commons Team**
- John Otridge
- Sima Pandya
- Todd Pihl
"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov