To develop the scientific evidence base that will lessen the burden of cancer in the United States and around the world.
Cancer Statistics

In 2016 there will be an estimated

1,700,000 new cancer cases

and

600,000 cancer deaths

- American Cancer Society 2016

Cancer remains the second most common cause of death in the U.S.

- Centers for Disease Control and Prevention 2015
Understanding Cancer

- **Precision medicine** will lead to **fundamental understanding** of the complex interplay between genetics, epigenetics, nutrition, environment and clinical presentation and **direct effective, evidence-based prevention and treatment**.
Cancer is a grand challenge

Requires:

Deep biological understanding
Advances in scientific methods
Advances in instrumentation
Advances in technology
Data and computation

Cancer Research and Care generate detailed data that is critical to create a learning health system for cancer
How do we solve problems in Cancer

- Support and incentives for team science, collaboration
- We need FAIR, open data
- Support open source, open science
- Support for rapid innovation
Cancer Moonshot

- Precision Medicine Initiative (PMI)
- National Strategic Computing Initiative (NSCI)

- Making data available: Genomic Data Commons
- Using the cloud: NCI Cloud Pilots
- Computation and data: DOE-NCI Pilots

- Audacious yet possible
- Investigate, explore, predict using real-world data!
Cancer Research Data Ecosystem – Cancer Moonshot BRP

Discovery

- Proteogenomics
- Imaging data
- Clinical trials

Well characterized research data sets

GDC

Research information donor

Patient engaged Research

- Clinical Research
- Observational studies

Cancer cohorts

Active research participation

Surveillance

Big Data

Implementation research

- EHR, Lab Data, Imaging,
- PROs, Smart Devices,
- Decision Support

Patient data

SEER

Learning from every cancer patient
The Cancer Genomic Data Commons (GDC) is an existing effort to standardize and simplify submission of genomic data to NCI and follow the principles of FAIR – Findable, Accessible, Attributable, Interoperable, Reusable, and Provide Recognition.

The GDC is part of the NIH Big Data to Knowledge (BD2K) initiative and an example of the NIH Commons.

Microattribution, nanopublications, tracking the use of data, annotation of data, use of algorithms, supports the data /software /metadata life cycle to provide credit and analyze impact of data, software, analytics, algorithm, curation and knowledge sharing.

Force11 white paper
https://www.force11.org/group/fairgroup/fairprinciples
NCI Genomic Data Commons

- The GDC went live on June 6, 2016 with approximately 4.1 PB of data.
- This includes: 2.6 PB of legacy data;
- and 1.5 PB of “harmonized” data.
- 577,878 files about 14,194 cases (patients), in 42 cancer types, across 29 primary sites.
- 10 major data types, ranging from Raw Sequencing Data, Raw Microarray Data, to Copy Number Variation, Simple Nucleotide Variation and Gene Expression.
- Data are derived from 17 different experimental strategies, with the major ones being RNA-Seq, WXS, WGS, miRNA-Seq, Genotyping Array and Expression Array.
- Foundation Medicine announced the release of 18,000 genomic profiles to the GDC at the Cancer Moonshot Summit.
## GDC Content

### Current
- TCGA: 11,353 cases
- TARGET: 3,178 cases

### Coming soon
- Foundation Medicine: 18,000 cases
- Cancer studies in dbGAP: ~4,000 cases

### Planned (1-3 years)
- NCI-MATCH: ~5,000 cases
- Clinical Trial Sequencing Program: ~3,000 cases
- Cancer Driver Discovery Program: ~5,000 cases
- Human Cancer Model Initiative: ~1,000 cases
- APOLLO – VA-DoD: ~8,000 cases

Total cases: ~58,000 cases
GDC Data Harmonization
Multiple data types and levels of processing

Exome-seq
1° processing
Genome alignment
2° processing
Mutations
3° processing
Oncogene vs. Tumor suppressor

Whole genome-seq
Genome alignment
Mutations + structural variants
Translocations

RNA-seq
Genome alignment
Digital gene expression
Relative RNA levels
Alternative splicing

Copy number
Data segmentation
Copy number calls
Gene amplification/deletion
PMI – Oncology, the GDC and the Cloud Pilots Goals

- **Support** precision medicine-focused clinical research
  - Enable researchers to deposit well-annotated (Interoperable) genomic data sets with the GDC
  - Provide a single source (and single dbGaP access request!) to **Find** and **Access** these data
  - Enable effective analysis and meta-analysis of these data without requiring local downloads – data **Reuse**
  - Understand **Contributions**, **Assess** value through usage, and give **Attribution** to all users
PMI – Oncology, the GDC and the Cloud Pilots Goals

- **Provide** a data integration platform to allow multiple data types, multi-scalar data, temporal data from cancer models and patients through **open APIs**
- Work with the Global Alliance for Genomics and Health (GA4GH) to **define** the next generation of **secure**, flexible, meaningful, interoperable, lightweight interfaces – open APIs
- Engage the cancer research community in evaluating the **open APIs** for ease of use and effectiveness
Questions?

Warren Kibbe, Ph.D.
Warren.kibbe@nih.gov
@wakibbe