Valentina Di Francesco
NHGRI Computational Genomics and Data Science Program
NHGRI Feb 2017 Council

The NHGRI Sandbox Group

- Lisa Brooks
- Jeff Kim
- Dan Gilchrist
- Carolyn Hutter
- **Kevin Lee**
- Samuel Moore
- Vivian Ota Wang
- Laura Rodriguez
- Heidi Sofia
- Jennifer Troyer
- **Chris Wellington**
- **Ken Wiley**

Outline

- Current challenges of genomic data sharing and analysis
- The Sandbox
- Interoperability with other genomic data commons
- Users
- Funding mechanism

Genomic Data Sharing and Analysis

Challenges

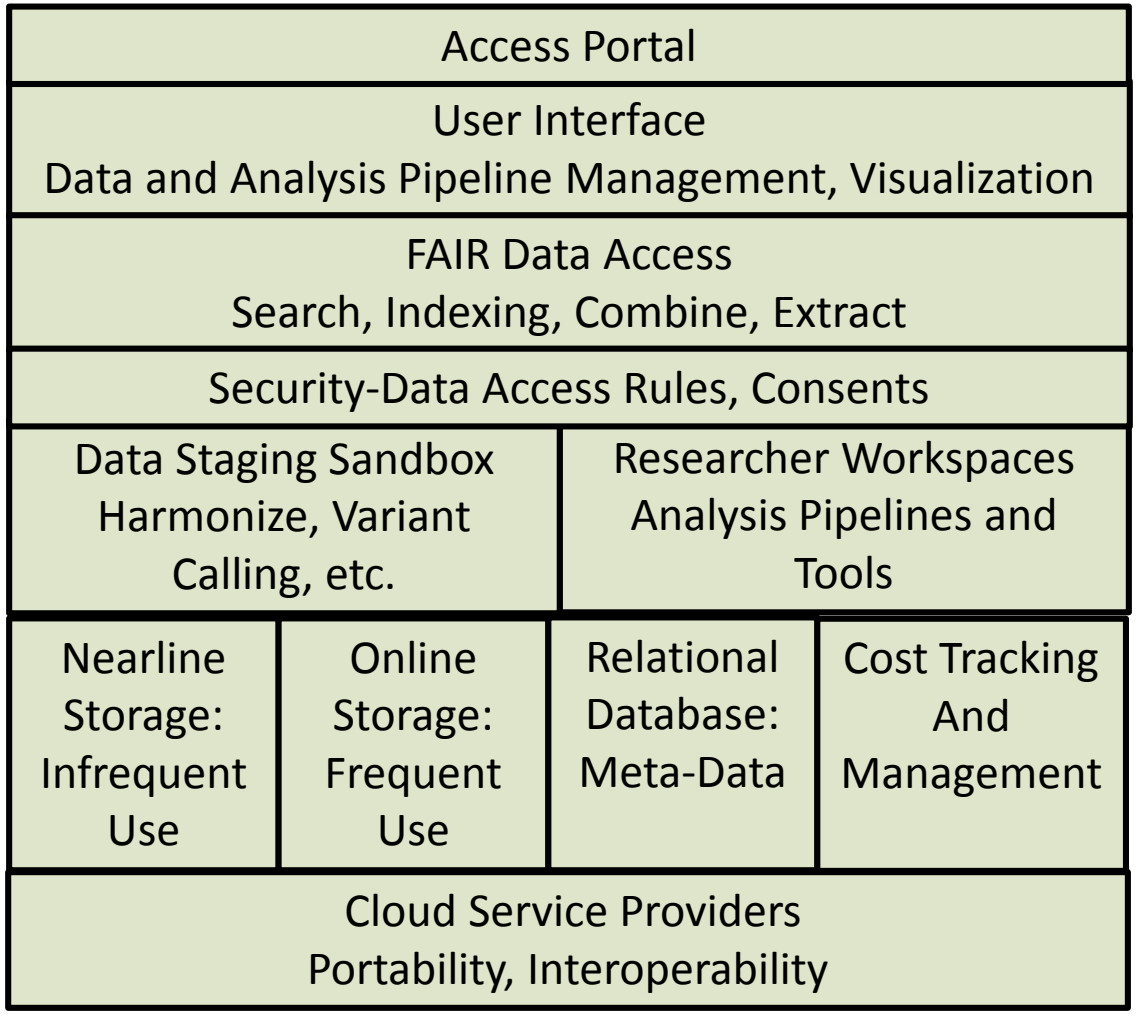
- Increasing sequencing capacity and decreasing sequencing costs lead to data management and analysis bottlenecks
- Researchers need scalable high-performance storage and computing infrastructure, and technical expertise.
- The current genomic data distribution and usage systems are inefficient and/or wasteful
  - Slow upload/download over the network
  - Multiple copies of same datasets in different local systems and cloud service providers
- Phenotypic data integration necessitates coordination
- NCBI/dbGaP is reducing its data archival role for the NIH

NHGRI Sandbox

A resource to democratize genomic data access, sharing and computing

- Leverages scalable cloud-based infrastructure to co-locate data storage and computing capacity with commonly used tools for analyzing and sharing data
- Provides access to unrestricted and controlled-access data and metadata from NHGRI funded programs – it is a Trusted Partner of dbGaP
- Performs metadata harmonization across NHGRI funded programs
- Makes available optimized, commonly used genomic analysis workflows and shares their results.
- Facilitates interoperability with other data resources and data commons

The Sandbox Architecture



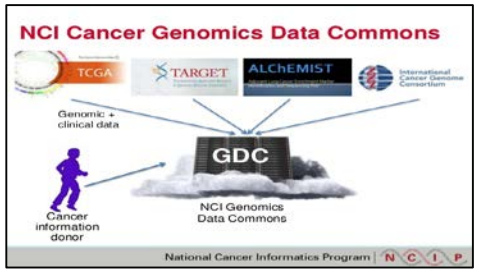
Courtesy of Vivien Bonazzi (NIH OD)

- NIH Notice for Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing Policy
- Data reside in a FISMA moderate, HIPAA compliant cloud system

NHGRI Sandbox

Federated Data Commons Model

NIH Data Commons



NIH HUMAN MICROBIOME PROJECT

TOPMed

All of Us™ The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU

Interoperability Tools

- Common user authentication system
- Shared APIs for data access and computing
- Adoption of FAIR Principles
- Docker containers
- Workflows management
- Digital objects catalogues and IDs
  ta standards and
  ologies

NHGRI Sandbox

ENCODE

emerge network
ELECTRONIC MEDICAL RECORDS AND GENOMICS

1000 Genomes

GSP

Sandbox's Users

User types

- Computational genomics scientists who are familiar with cloud or high-performance computing.
- Researchers who do not have extensive coding experience and may want to use simpler tools and web interfaces provided by the Sandbox.

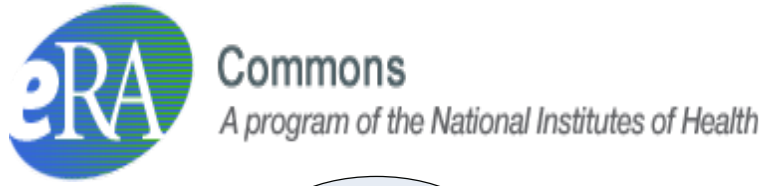
Individual users

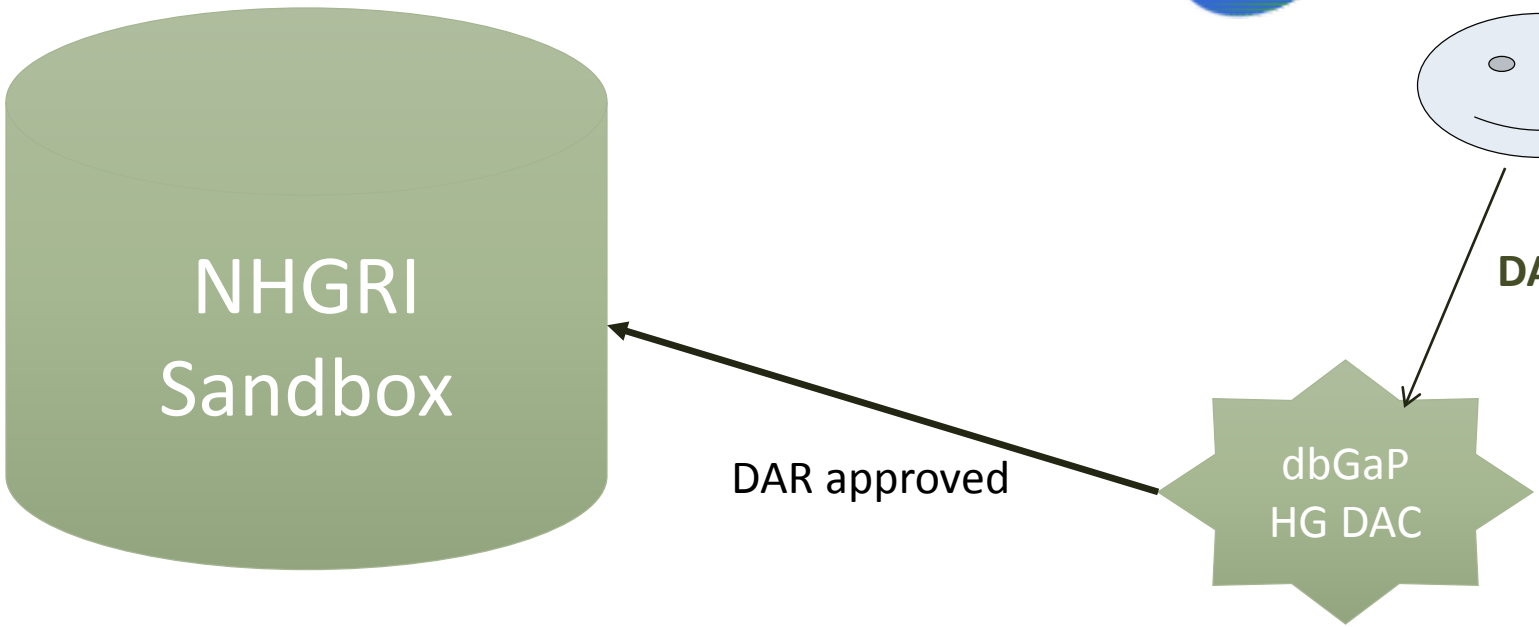
- Are able to upload and compute on the Sandbox's and their own data combined.
- Have access to data without download to local storage and analysis systems.
- Have access to commonly used analysis tools and workflows.
- Have a workspace for new tool development and sharing.

Controlled Access Data

No transfer of data from dbGaP to another system

User authentication





User authorization

Potential Costs to Individual Users

Data Storage

- $350-$450 x TB/year depending on data access speed

Computing

- WGS variant calling using BWA-Mem & GATK 3.4 on 60GB input @ 30X cov. - $270/sample AWS on demand or $30/sample spot instance
- RNA-Seq using Tuxedo tool suite on 7GB @ 30 cov. $16/sample AWS on demand or $2/sample spot instance

Data Egress

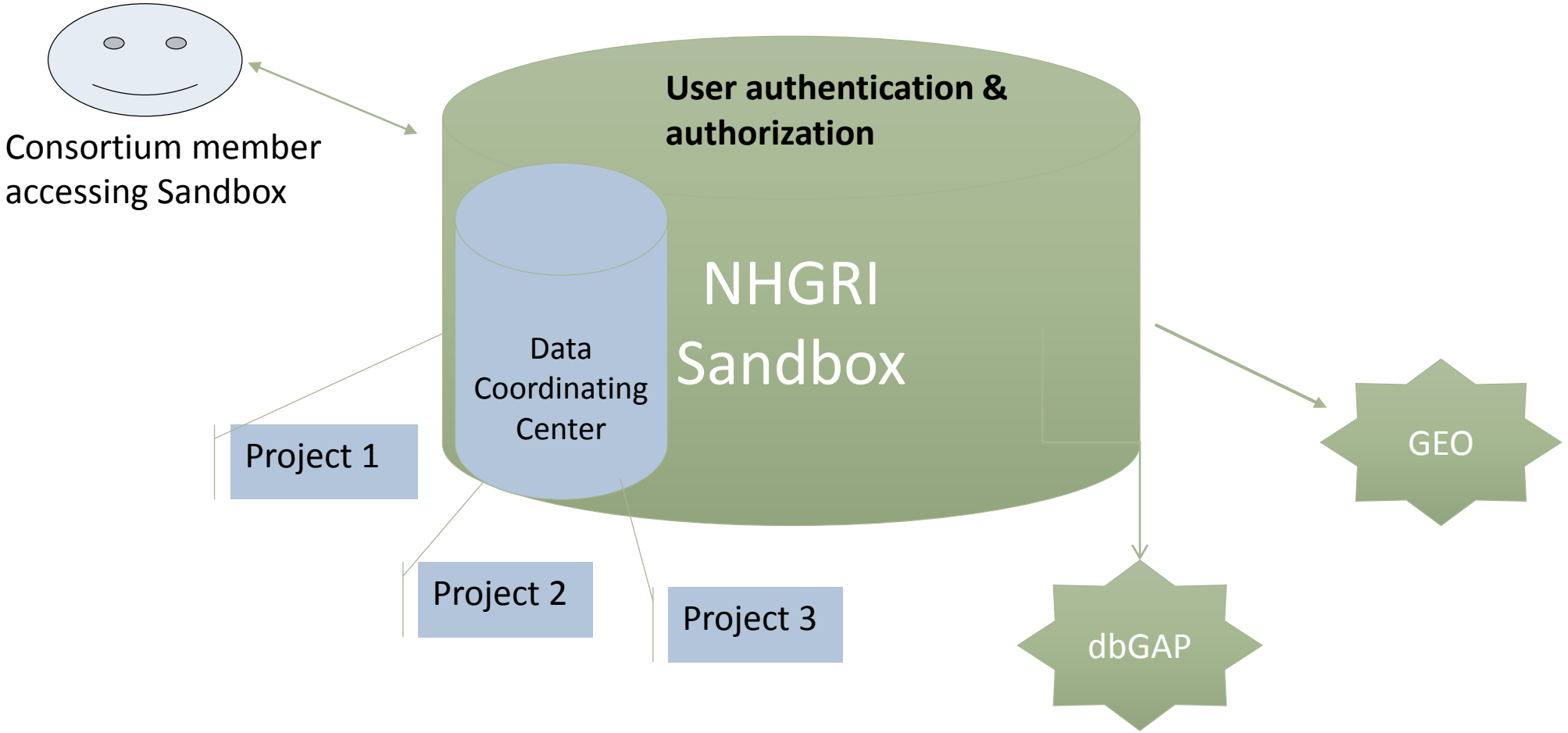
NHGRI Sandbox

NHGRI Consortia and the Sandbox

NHGRI consortium members

- Have access to cloud resources where they can deploy new data sets and workflows.
- Have a common infrastructure to share and compute on consortium data.
- Take advantage of interoperability tools and data submission services to other data commons.

NHGRI Programs and the Sandbox



Funding Mechanism

Contract

- Required to establish a trusted partnership with dbGaP
- Deliverables, special reporting requirements, close collaboration and oversight by NHGRI staff

Funding period: 7 years

Timeline

- 1 award expected in Summer 2018
- Kick off in Spring 2019

NHGRI Sandbox