# Central analysis server

Mark DePristo, Lisa Brooks, Pearl O'Rourke, Carlos Bustamante, David Altshuler

NIH Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects, June 2012

# What are we trying to accomplish?

To make it possible for researchers to answer scientific questions about the relationships between inherited DNA variation and human phenotypes

- For example:
  - Given a disease of interest, what set of human genes harbor DNA variation that is robustly associated with risk of disease or somatic mutation in cancer?
  - Given a gene of interest, what phenotypes (if any) are associated with inherited and somatic DNA variation?

# Why can we not achieve our goal today?

- Until recently, there existed fundamental barriers to integrated analysis of genotype and phenotype

- Data about each disease was incomplete and of inadequate scale
  - Collected and analyzed in silos, one phenotype and one sample set at a time, without clear routes to access
  - Unmeasured confounders due to different technical platforms

- Diverse analytical methods were developed but
  - It is difficult for methods devs and analysts to access the data
  - It is difficult for data holders to run most methods
  - Many computational methods were not instantiated in software of sufficient quality to perform (let alone automate) analysis

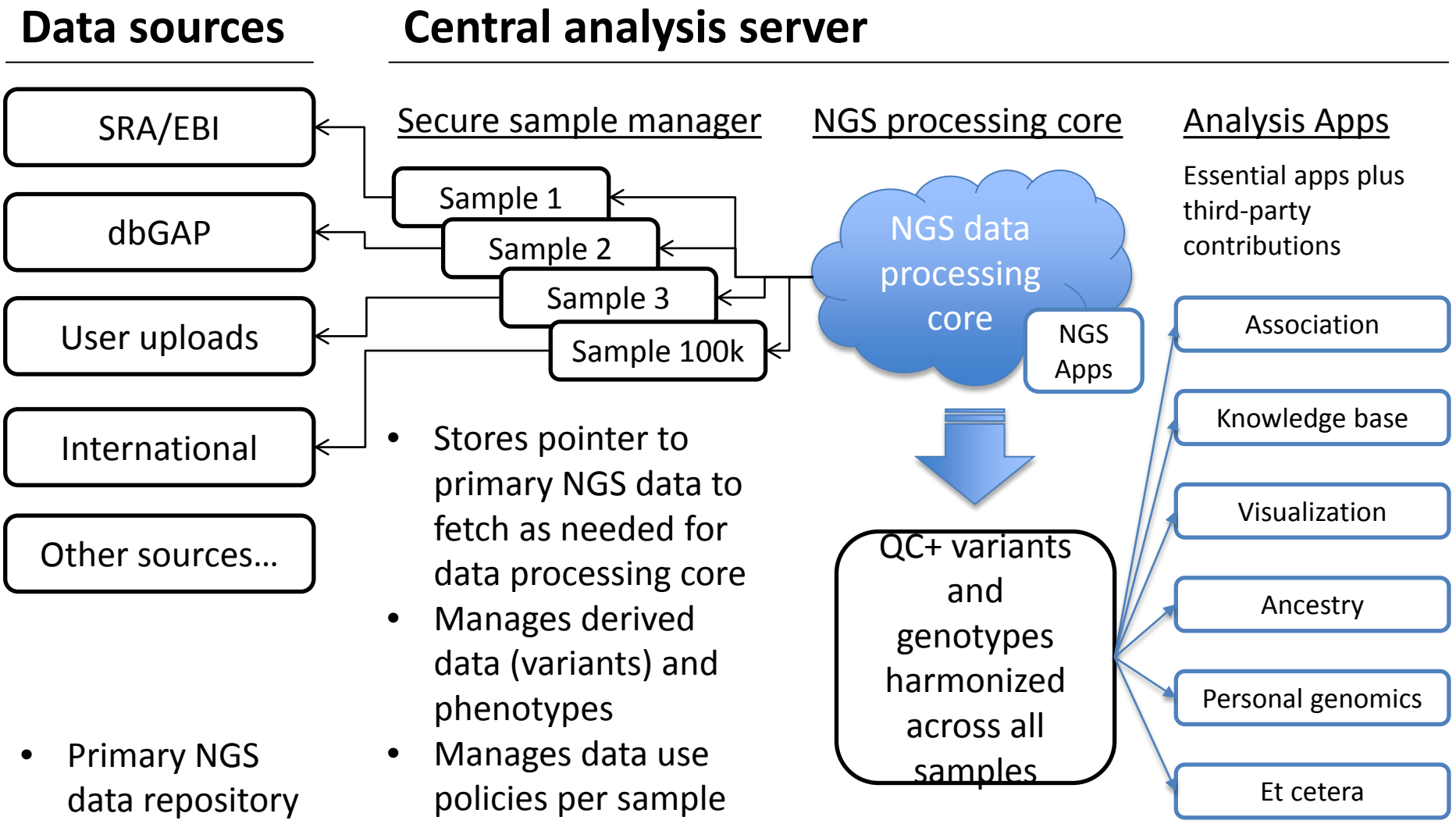
# These barriers have fallen

- Because of next-generation sequencing
  - Rich data type with intrinsic QC properties
  - >50,000 genome sequences already generated
- Because dbGAP provides a route to access
- Because of newer analysis tools
  - That can integrate and harmonize data collected at different sites and with different platforms
  - That can perform data processing and association analysis in an automated manner

# We could seize this opportunity to build a central analysis server

- To aggregate in a single location available data on human DNA sequence and phenotype
- To provide a state-of-the-art computational environment and analysis tools to manage, process, and analyze the data for phenotypic association
- To managing security, data use, and user access to ensure that each dataset is used only in manners allowed by the original informed consent and data use agreements

# Yes, that's nice but how would this work?



# A few key features of the server

- The main development need is to build a computational Platform to
  - Coordinate sample NGS data and phenotypes
  - Understand and enforce data use policies per sample and per user
  - Execute apps at the scale of 100,000s of samples
- The platform must be continuously updating and evolving
  - Must be able to upload new Apps for users (after vetting)
  - Must regularly update analysis of all samples with the best methods
  - Must continuously monitor data sources for new samples
- Need apps for variation discovery in 100Ks of samples
  - Product is harmonized, error corrected polymorphic sites across samples
  - Genotypes and their likelihoods for each sample at every site
  - Joint error modeling across samples to remove errors

# The Platform must understand and enforce data use restrictions per sample and user

## Today

- Sign a piece of paper
- Download raw data from dbGAP
- "On honor" to follow use and users policy

## In the Platform

- Represents data use restriction for each sample
  - Ex: IRB consent, NIH
- Enforces these policies per user
  - Ex: P. Investigator only
- Restricts analyses to only those allowed per user across samples
  - Ex: Autism researcher looking for rare variation in Autism case samples

**Example data use policies**

1000 Genomes samples

Users: freely available
Uses: no restrictions

NIMH Autism samples

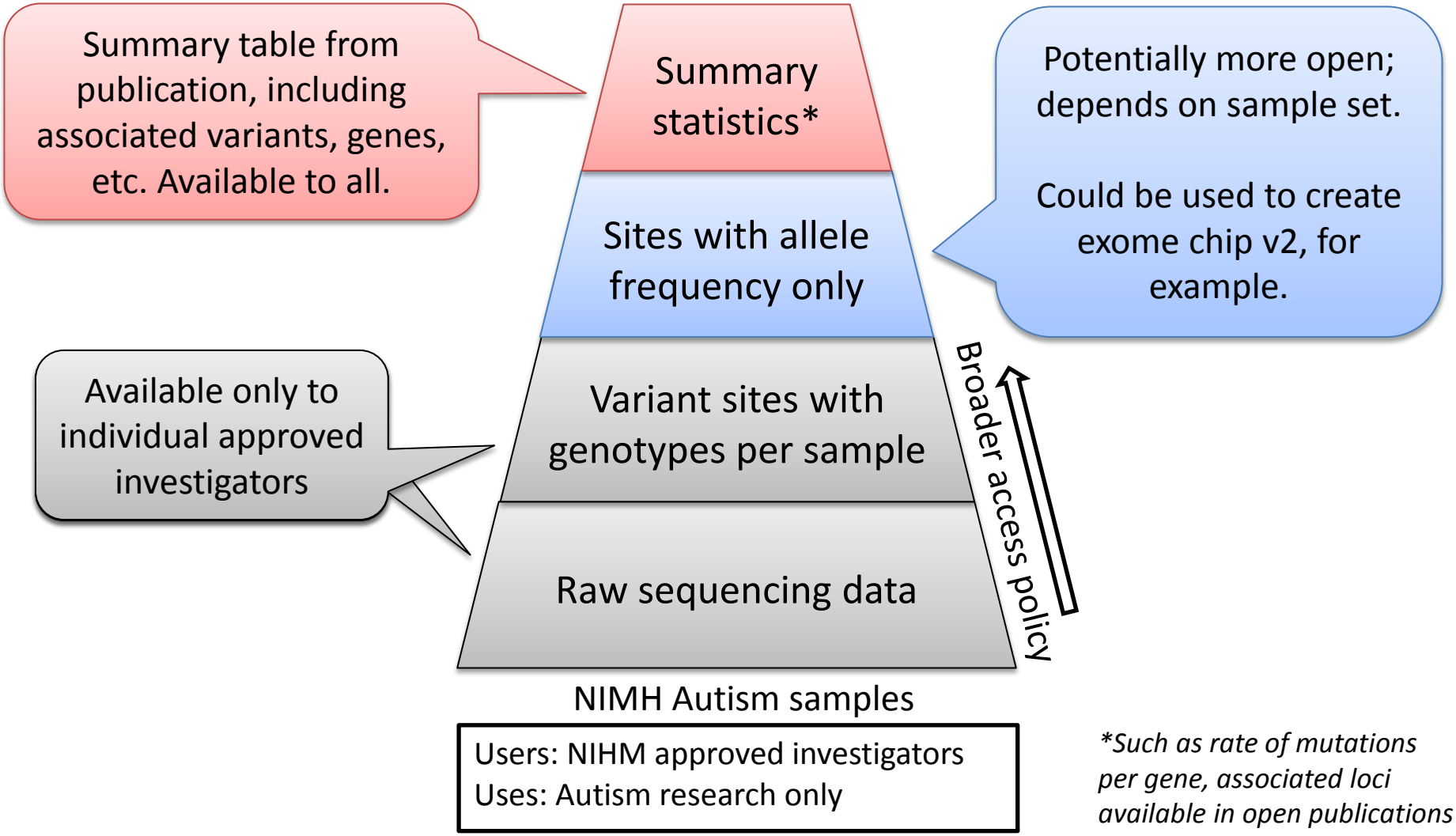
Users: NIHM approved investigators
Uses: Autism research only

NIMH common controls

Users: NIHM approved investigators
Uses: any condition or trait

Goal: to force compliance with existing data use policy, whatever they may be

# The Platform must track data use policies at multiple levels of data detail



# The server will support analysis of samples across the sharing spectrum

**1000 Genomes sample**

- Low-level data is freely available
- The server would allow anyone to operate on them in any way possible given the server's apps
- Can we freely merged with other data sets

**Single use sample**

- E.g., Schizophrenia by non-commercial users
- Investigators studying Schizophrenia could still use the server data for
  - Analysis infrastructure
  - Access to shared controls
- But no studies of other phenotypes could see this sample or benefit from these data

Most samples fall somewhere in the middle

# (Some) advantages of a central server

- Enhances the value of large collections of shareable data by streamlining access
  - E.g., the common NIMH controls and the 1000 Genomes Project
- Enjoys a strong network effect to drive adoption
  - Access to cutting-edge Apps for data processing and analysis
  - Easy access to comparator datasets
- Knowledge base of variation, phenotype association, and supporting NGS data provides a natural interface to:
  - Biologists – who can determine what variants (and phenotype associations) are known in their gene of interest
  - Pharma / biotech – who can explore the impact of rare (e.g., LoF) variation across genes and disease
  - Geneticists – who can easily incorporate large-scale high-quality control data into their association studies

# Conclusions

- Central server would provide a sustainable infrastructure for integrative genetic analyses
- Computational and ELSI challenges are real but manageable
- A successful server would be immensely valuable
  - Should build multiple servers to encourage innovation and diversity
  - Servers specific to disease areas, e.g., cancer?

Serving our larger ecosystem

Biologist / Pharma / Biotech

#server It's a comprehensive knowledge base for trait genetics

Geneticist

#server It's the largest sample size for my disease study

Statistician

#server It's data for my models!

Method developer

#server So easy to share my method with the community

ELSI

#server It understands the rules that protect the people behind the samples

# Appendix

# ELSI consideration for the server

- The central server would likely be considered a research protocol and would need to be reviewed and approved by an IRB
- No sharing of individual-level data from the system

- The IRB protocol should include details on the “business rules” such as:
- Requirements for submission of data to the server
  - All data will have been generated from tissue obtained under an IRB-approved informed consent form (ICF) process
  - Server Data Managers would confirm agreed-upon standards before the server accepted a dataset, such as that any HIPAA identifiers were removed, and would obtain the data use conditions from the DAC that approves use of the data set, or from the originating institution.

- Requirements on use of data
  - High interest analyses will be pre-computed with results made available to all users
  - Requests for custom analyses will be submitted in a uniform manner that allows automated systems to confirm that requests conform to data use conditions.

# Proof-of-concept data processing core

- We don't have general platform with user management and security
- We do have at BI practical, highly scalable infrastructure for processing NGS data to very high quality
- We performed joint data processing and error modeling of chr1 of ~16K exomes
  - Samples from 1000 Genomes, Autism, Diabetes, Schizophrenia, and the Exome Sequencing Project
  - Data generated at BI, Sanger, BCM, BGI, Germany, among others over last 3 years with multiple capture technologies
- Machine learning approach to find errors among all samples simultaneously
- Product is all polymorphic sites across all samples and the genotypes (and likelihoods) of each sample at each site
- The BI/GATK data processing tools would be among the first installed in the Platform

Computational requirements to analyze chr1 of 16K exomes

| | |
|---|---|
| Samples | 16,373 |
| Per-sample BAM processing | 16K CPU days |
| Analytic BAM storage | ~18 MB |
| Joint calling | 120 CPU days |
| | |
| Polymorphic sites found | 196K |
| Likely SNP artifact removed (% LoF) | 43K |
| Likely indel artifacts removed (% LoF) | 2.2K |
| Chr1 VCF | 93 GB |