National Cancer Institute

NCI Cancer Genomics Cloud Pilots (and Genomic Data Commons)

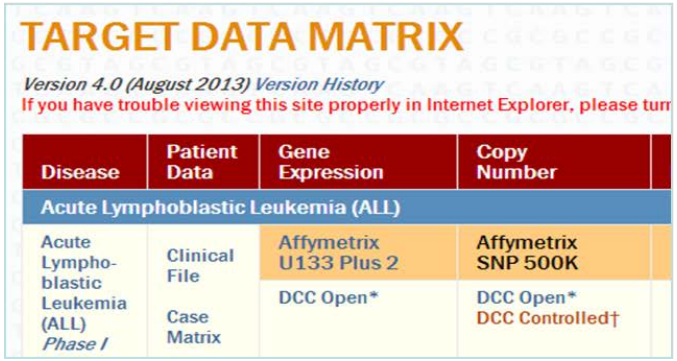
Tanja Davidsen, Ph.D.
Center for Biomedical Informatics and Information Technology (CBIIT)
National Cancer Institute

May 12, 2015

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Center For Cancer Genomics (CCG) Genomics Data Commons (GDC)

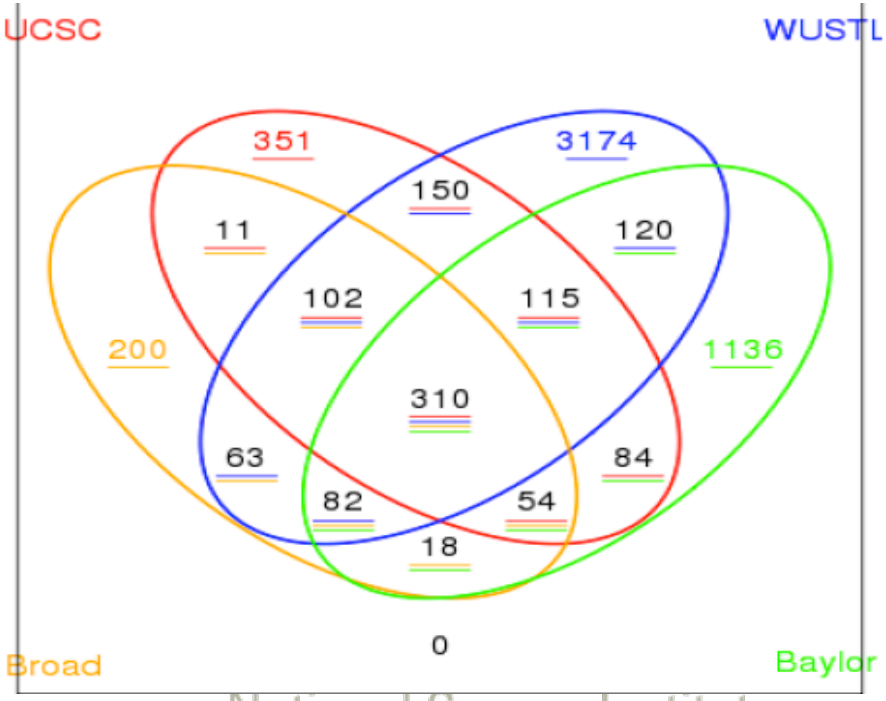
- Goal to unify fragmentary repositories at NCI
- TCGA, TARGET and CGCI have their own data repositories (DCCs)
- Sequencing data: BAM files at CGhub while VCF/MAF files at DCC



Genomics Data Commons (GDC)

- Harmonize diverse standards
- BAMs aligned to various references
- Mutations are called by various tools

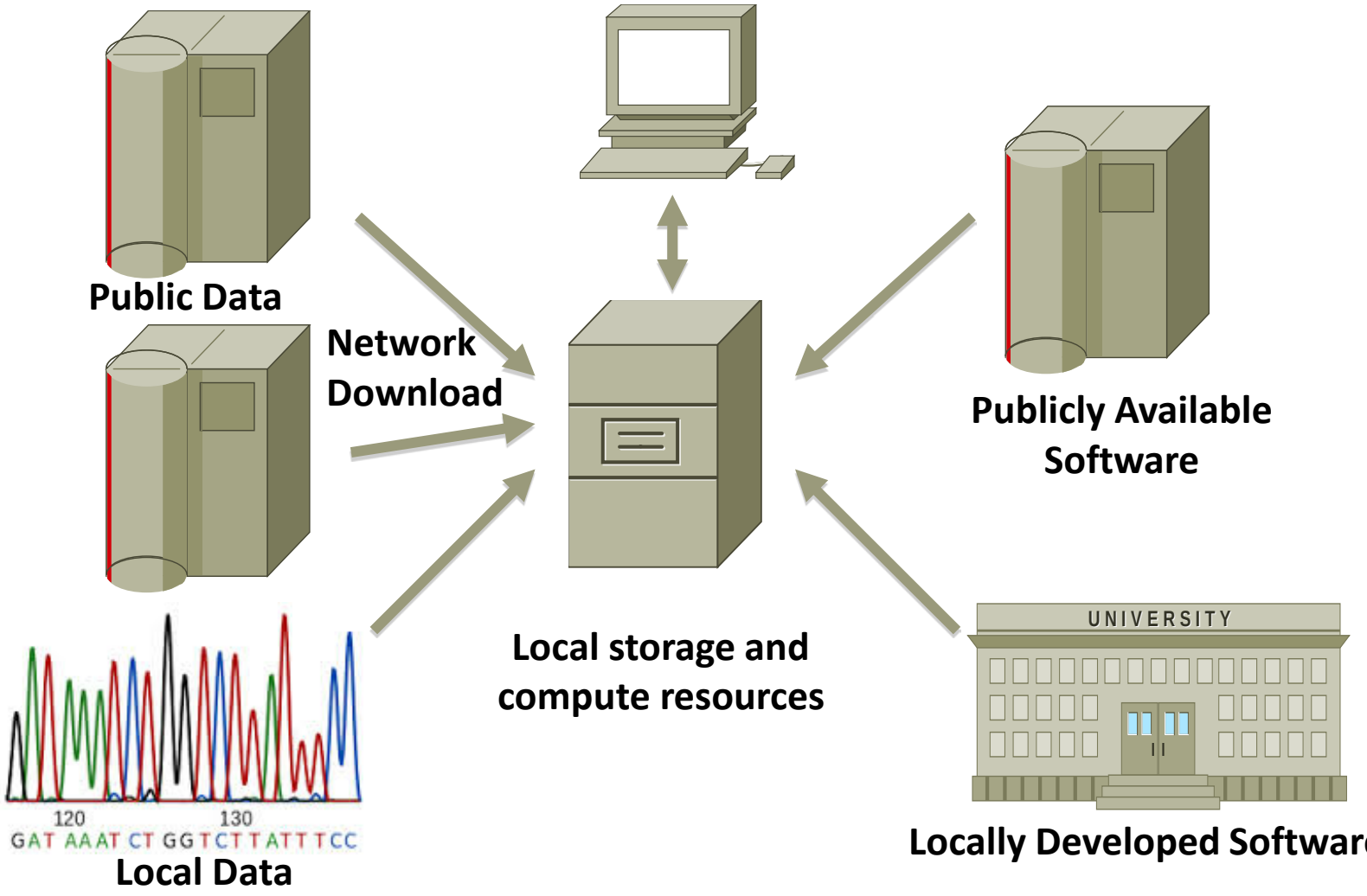
NCBI-human-build36
NCBI36_BCCAGSC_variant
NCBI36_BCM_variant
NCBI36_WUGSC_variant
HG18
HG18_Broad_variant
GRCh37
GRCh37-lite
GRCh37_BI_Variant
GRCh37-lite-+-HPV_Redux-build
HG19
HG19_Broad_variant



Genomics Data Commons (GDC)

- University of Chicago, PI: Dr. Robert Grossman
- Go live date: Late Spring 2016
- Not a commercial cloud: Free to download data

Standard Model of Computational Analysis

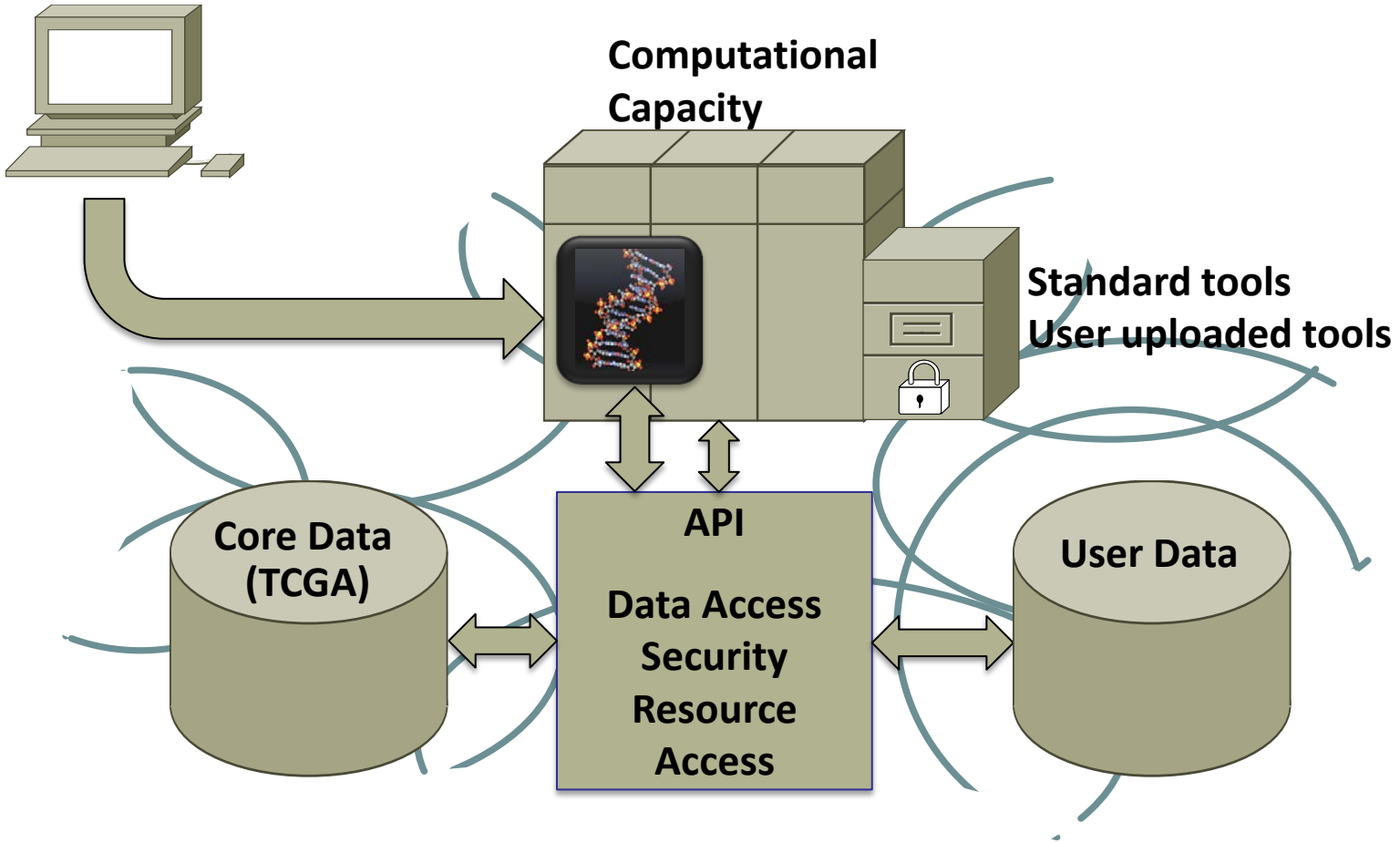


National Cancer Institute

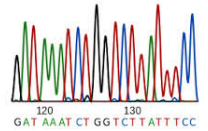
Limitations of the standard model for large data

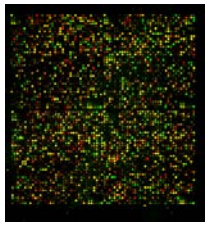
- Assuming the 2.5 PB TCGA data set
- Storage and Data Protection cost approximately $2,000,000 per year
- Downloading TCGA data at 10 Gb/sec = 23 days
- Only large institutions have the ability to utilize this data
- These datatypes will continue to grow

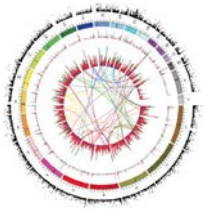
Co-located Compute & Data

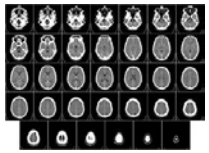


The Cloud Pilots in Context









Data Coordinating Center

QA/QC Validation

Aggregation

Authoritative NCI Reference Data Set

NCI Clouds

High Performance Computing

Analysis

Search/Retrieve

Download

NCI Genomic Data Commons

National Cancer Institute

Project Structure

Effort to democratize access to NCI genomics data

Managed through CBIIT in partnership with the Center for Cancer Genomics (CCG)

- Coordinating with the Genomic Data Commons (GDC)

Three contracts awarded to

- Broad Institute
- Institute for Systems Biology
- Seven Bridges Genomics

Period of performance: Sept 2014 – Sept 2016

- https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots
- Anticipated launch date: January 2016

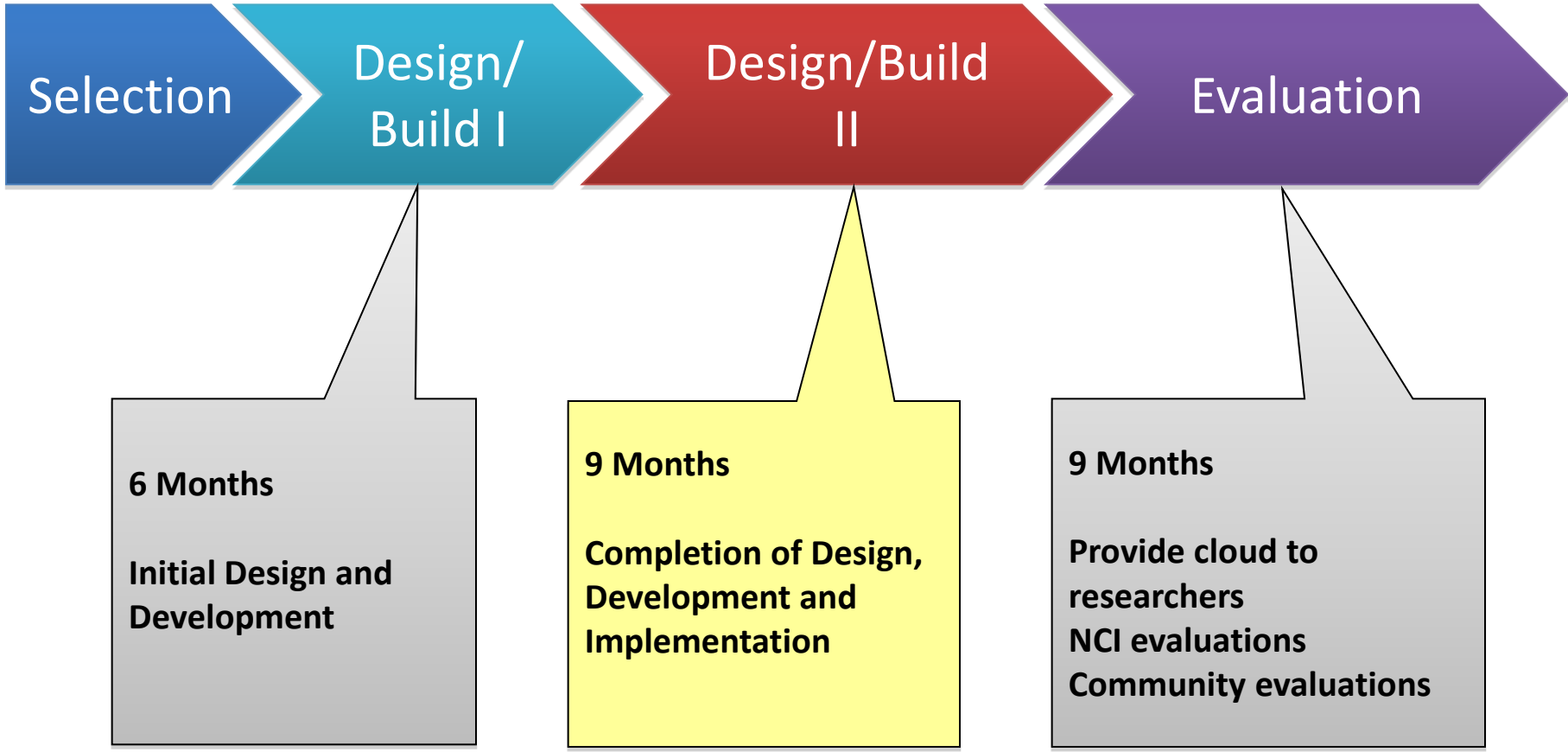
Considerations

- Design
  - Designs must be released under a non-viral, open source license
- Extensibility
  - Initial clouds will focus on a set of “core datatypes”
  - Extend to additional datatypes without major refactoring of the existing system
- Sustainability
  - Cost assessments for operating at current scale and at 10/100 fold increases in storage, compute and usage
- Security
  - FISMA moderate system, FedRAMP certified cloud provider, Trusted Partnership
  - Open v/s Controlled access data

Research and Technical Objectives

- Core Data
  - All three awardees will host a common core data set from TCGA
    - DNA-Seq binary alignment (BAM) files
    - RNA-Seq FASTQ and BAM files
    - SNP array (.cel) files
    - Somatic and germline mutation calls for each sample (.vcf, .maf)
    - Clinical data
  - Each awardee will include at least one additional TCGA data set
    - Broad: validation BAMs, miRNAseq, and methyl-seq
    - ISB: miRNAseq, and all L3 data (mRNA/miRNA expression, copy-number, DNA methylation, protein RPPA)
    - Seven Bridges: whole genome and exome DNA-Seq FASTQ, miRNAseq data, and methyl-seq

Project Schedule and Deliverables



Common to all three Cloud Pilots

- Core datasets
- Use Cases
  - Running preloaded pipeline on TCGA data
  - Uploading and processing user data
  - Uploading and running custom algorithms
  - Serve both biologists and bioinformaticians
- Workflow Language
  - Common Workflow Language (CWL) is being considered
- Docker containers
  - For improved portability and reproducibility
- Using emerging GA4GH standards
- Authorization and Authentication process

Broad Cloud Pilot

- PI: Gad Getz
- Collaborators: University California Berkeley, University California Santa Cruz
- Cloud Platform: Google
- Unique Technologies Used: ADAM/Spark
- Tools Incorporated: Firehose
- Cloud Pilot Website: http://firecloud.org

FireCloud

TCGA data

Your Data **Data**

Tools

Your Tools

ReCapSeg MuTect GISTIC PathSeq RNASeqQC SegSeq Gistic HapSeg MutSig2CV MutSigCVInterVar dRanger ABSOLUTE ContEst coMut InVexChainFinder MutSig1 MutSigCV RNASeQC CapSeg Oncotator

User or team Workspace

- Securely tracks and manages data, metadata, tools, job execution and results
- Captures provenance for each run (method versions, timestamps, input and output files)

Your Workflow

Workflows

Users

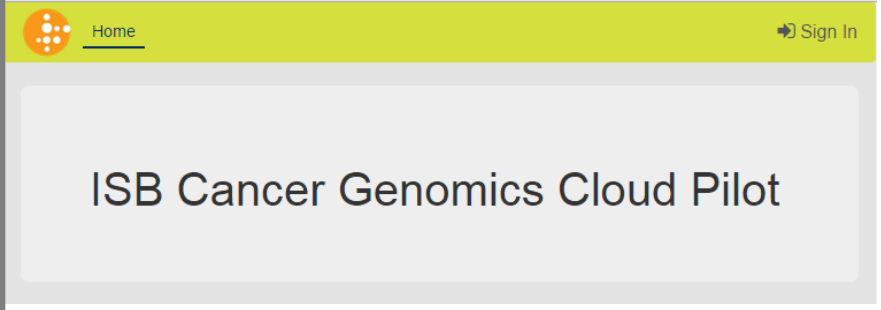
ProductionManager ProjectManager PI Group Lab ToolDeveloper Clinician Analyst

The FireCloud Platform

- Scaleable and elastic compute on Google Cloud Platform
- Versatile job executors:
  Docker, ADAM/Spark, Google Cloud Dataflow
- Data can be stored in scalable distributed 'stores':
  ReadStore, VariantStore
- Available Spring 2016
- GA4GH compliant data stores
- For more updates and information, leave your contact information at: *firecloud.org*

Institute for Systems Biology (ISB) Cloud Pilot

- PI: Ilya Shmulevich
- Collaborators: Google, SRA International
- Cloud Platform: Google
- Unique Technologies Used: Google Genomics Platform
- Tools Incorporated: Regulome explorer, Gene Spot
  - Focus on interactive data visualization, exploration and analysis
- Cloud Pilot Website: http://cgc.systemsbiology.net/



Interactive tools

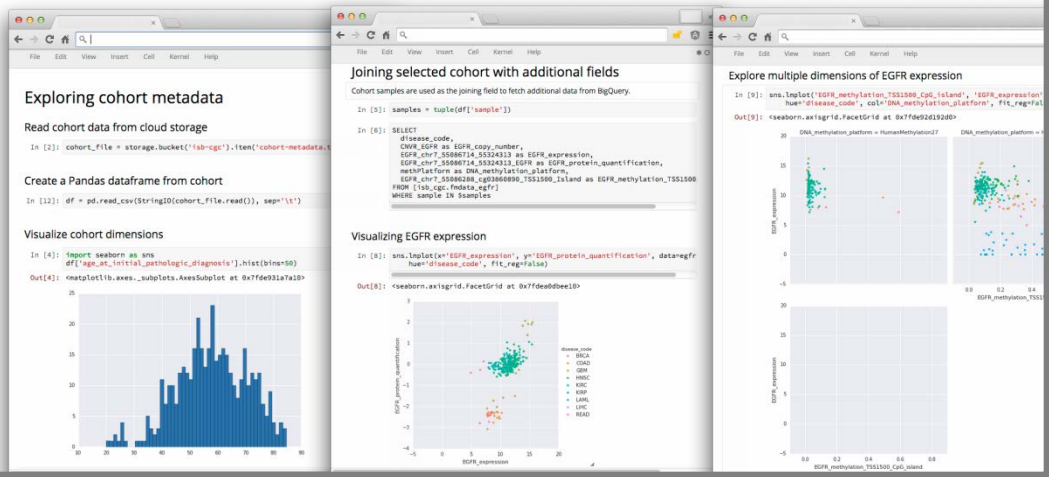
- explore all tumors or a subset
- define custom “cohorts”
- focus on specific molecular data types or platforms

Programmatic access

- REST APIs for Cloud Storage
- SQL-like queries for BigQuery
- GA4GH API for Google Genomics

Tutorials

- IPython notebooks
- RStudio (Rmd) files



Seven Bridges Genomics Cloud Pilot

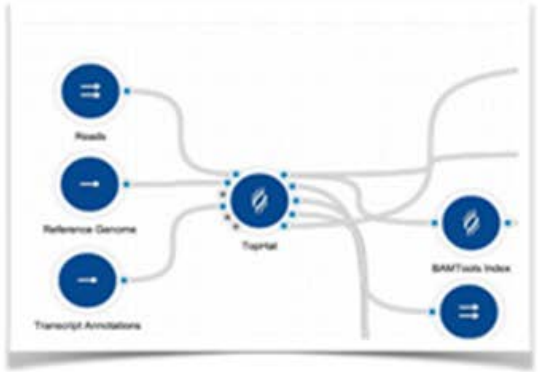
- PI: Deniz Kural
- Collaborators: None
- Cloud Platform: Amazon Web Services
- Unique Technologies Used: SBG platform
- Tools Incorporated: > 30 public pipelines
  - https://igor.sbgenomics.com/lab/public/pipelines/
- Cloud Pilot Website: https://www.sbgenomics.com/cancer-genomics-cloud/

CANCER GENOMICS CLOUD
SEVEN BRIDGES

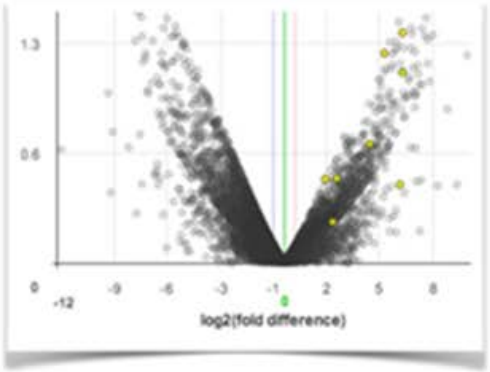
Run your tools on TCGA data. No waiting.

```
import os
from sbgsdk import define, Process


class SamtoolsSamToBam(define.Wrapper):
    class Inputs(define.Inputs):
        inp = define.input(required=True)

    class Outputs(define.Outputs):
        out = define.output()

    def execute(self):
        out_file_name = change_ext(self.inputs.inp,
        Process('samtools', 'view', '-S', '-b', sel
```

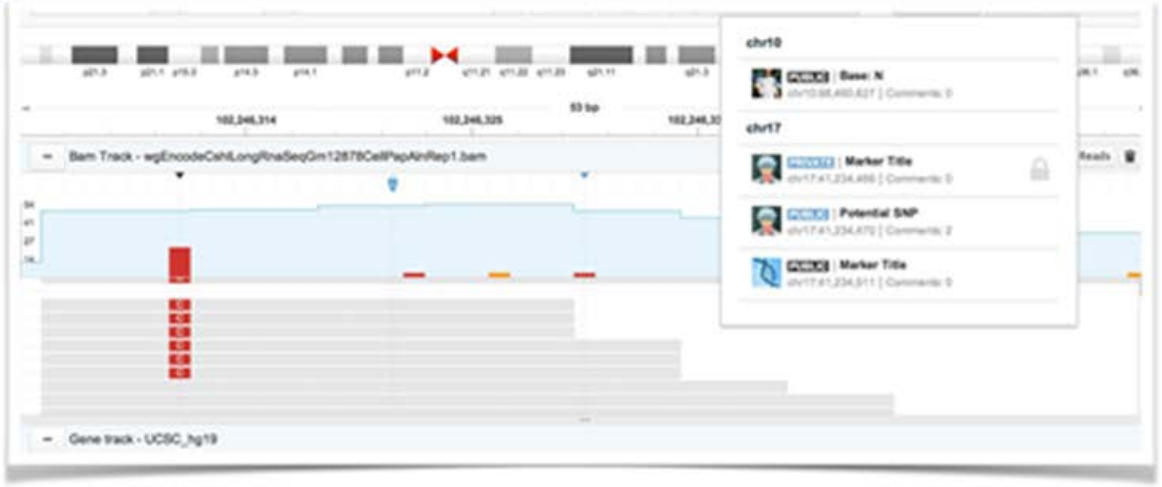
Wrap your tools



Mix and match with others …and TCGA data



Analyze



Collaborate and discuss what you discover with colleagues

Pre-register at http://www.cancergenomicscloud.org/

Cloud Pilot Workshop

- Additional details
- Today, 4-5 and 5-6pm
- Natcher Auditorium

Cancer Genomics Project Teams

CGC Pilot Team Principal Investigators

- **Gad Getz, Ph.D** - Broad Institute - http://firecloud.org
- **Ilya Shmulevich, Ph.D** - ISB - http://cgc.systemsbiology.net/
- **Deniz Kural, Ph.D** - Seven Bridges - https://www.sbgenomics.com/cancer-genomics-cloud/

NCI Project Officer & CORs

- Anthony Kerlavage, Ph.D – Chief Project Officer
- Juli Klemm, Ph.D – COR, Broad Institute
- Tanja Davidsen, Ph.D – COR, Institute for Systems Biology
- Ishwar Chandramouliswaran, MS, MBA – COR, Seven Bridges Genomics

GDC Principal Investigator

- Robert Grossman, Ph.D - University of Chicago

Center for Cancer Genomics Partners

- JC Zenklusen, Ph.D
- Daniela Gerhard, Ph.D
- Zhining Wang, Ph.D
- Liming Yang, Ph.D
- Martin Ferguson, Ph.D

NCI Leadership Team

- Warren Kibbe, Ph.D
- Lou Staudt, Ph.D
- Steven Chanock, Ph. D
- George Komatsoulis, Ph.D


DEPARTMENT OF HEALTH & HUMAN SERVICES · USA
NIH
NATIONAL® CANCER INSTITUTE