SCALABLE, COLLABORATIVE, REPRODUCIBLE, AND EXTENSIBLE ANALYSIS OF TCGA DATA IN THE CLOUD

Brandi Davis-Dusenbery, PhD

CBIIT Speaker Series

January 6, 2016

CANCER GENOMICS CLOUD

SEVEN BRIDGES

AGENDA

- Motivation
- Guiding Principles
- Case study

CANCER GENOMICS CLOUD
SEVEN BRIDGES

TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY …

More than 11,000 cases representing 33 cancer types

CANCER GENOMICS CLOUD
SEVEN BRIDGES

TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY …

Primary Tumor

Metastatic

…

Blood Derived Normal

Solid Tissue Normal

…

multiple Samples per Case

Genomic

Proteomic

…

Transcriptomic

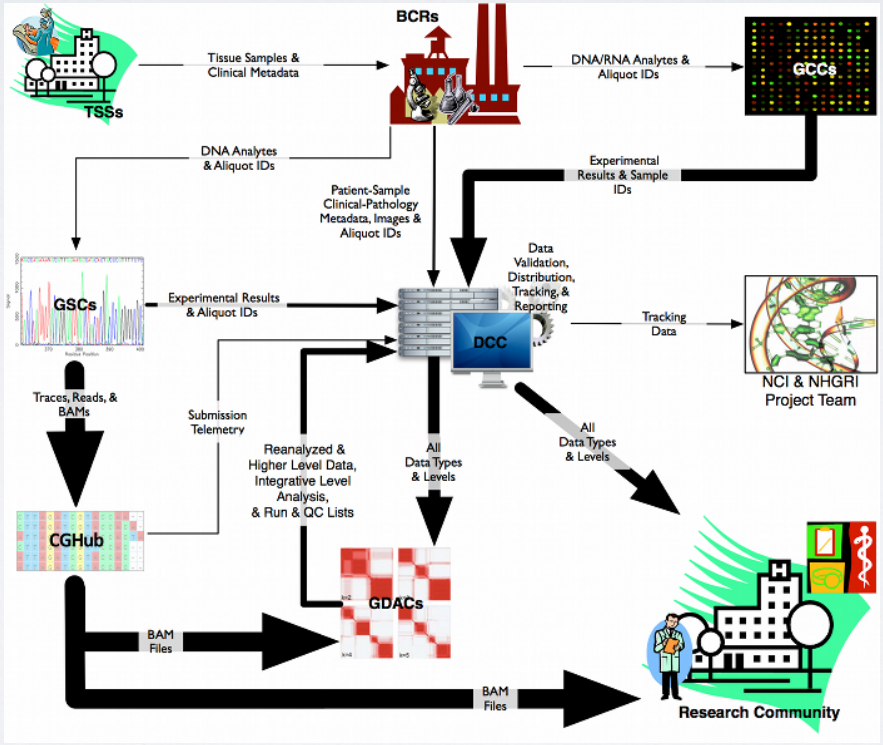
Epigenomic

…

multiple Analyses per Sample

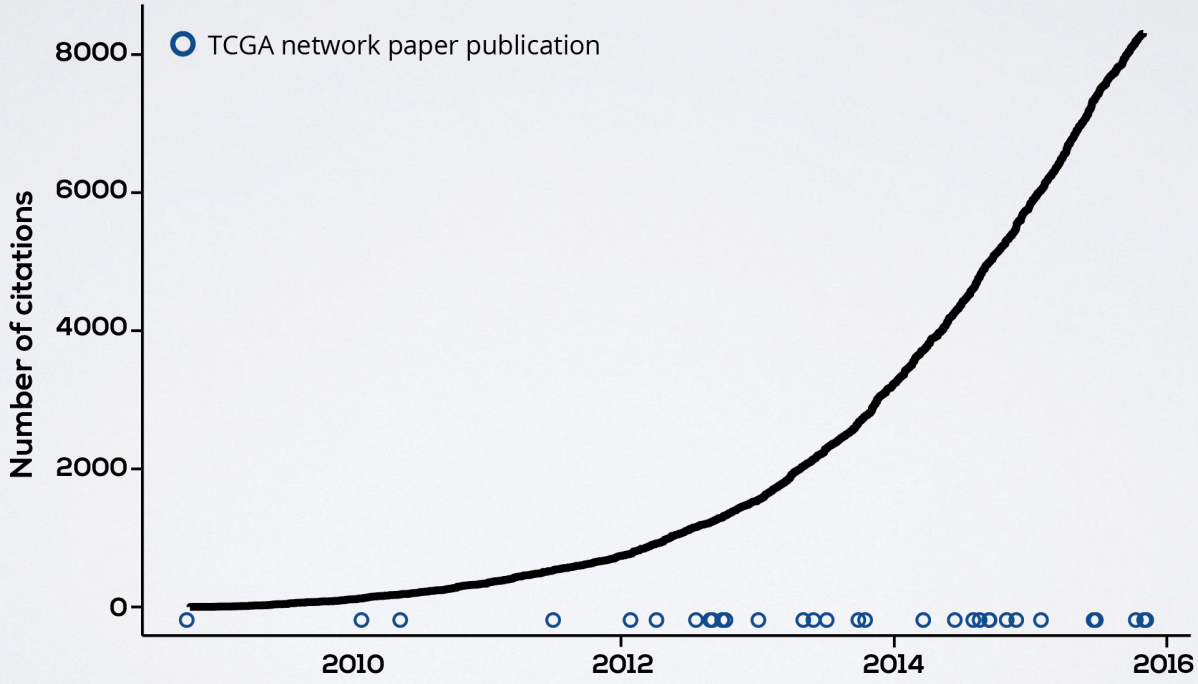
CANCER GENOMICS CLOUD
SEVEN BRIDGES

… MADE POSSIBLE BY THOUSANDS OF RESEARCHERS …



CANCER GENOMICS CLOUD
SEVEN BRIDGES

… WITH FAR REACHING IMPACT.

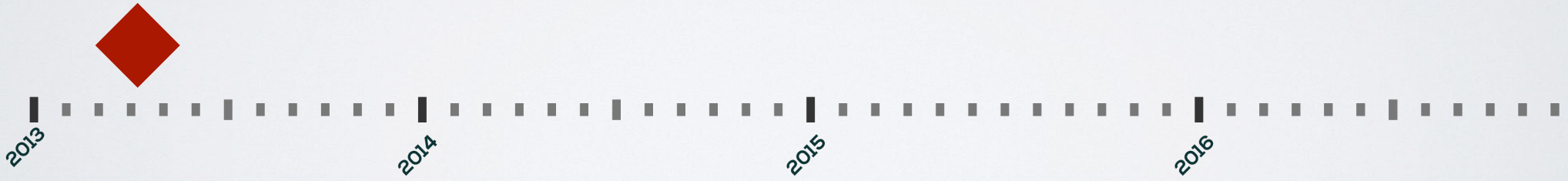


CANCER GENOMICS CLOUD
SEVEN BRIDGES

However, as the amount and diversity of data increases,
it becomes more difficult to learn from them.

3 YEARS IN THE MAKING…

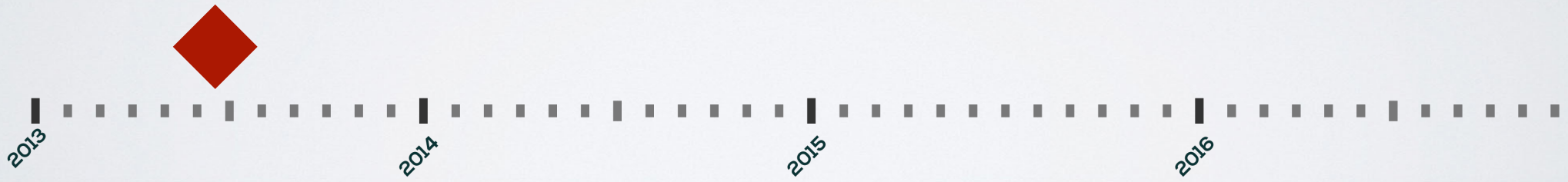
April 2013: Recognizing these challenges, Dr. Harold Varmus & colleagues issue letter proposing creation of public "cancer knowledge clouds" and seeking input from the research community on data storage and compute challenges.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

3 YEARS IN THE MAKING…

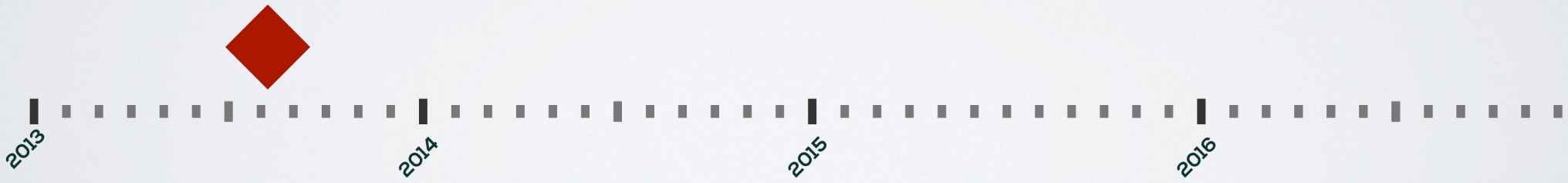
June 2013: The Cancer Genomics Cloud Pilot concept, presented by Dr. George Komatsoulis receives unanimous approval at a joint meeting of the NCI Board of Scientific Advisors and the National Cancer Advisory Board.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

3 YEARS IN THE MAKING…

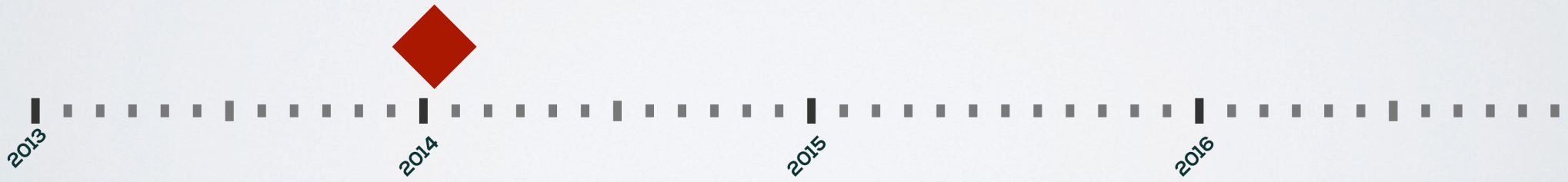
August 2013: Community feedback regarding capabilities and critical use cases collected via IdeaScale site & Sources Sought notice



CANCER GENOMICS CLOUD
SEVEN BRIDGES

3 YEARS IN THE MAKING…

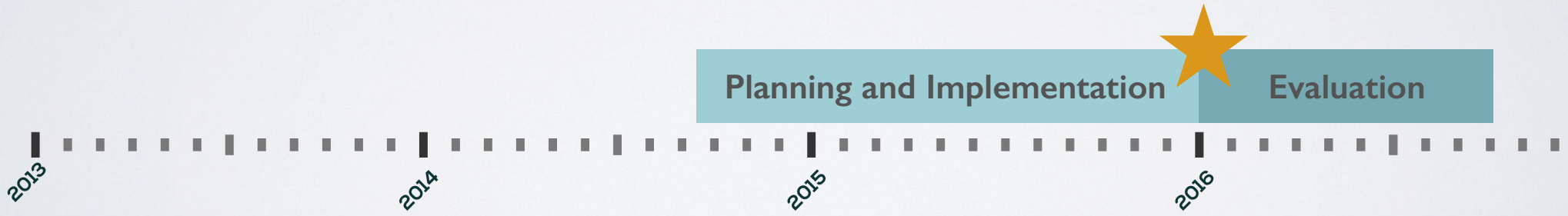
January 2014: Broad Agency Announcement issued to support development of the pilots



CANCER GENOMICS CLOUD
SEVEN BRIDGES

3 YEARS IN THE MAKING…

September 2014: The Broad Institute, Institute for Systems Biology and Seven Bridges awarded two year contracts to build pilot systems.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

GUIDING PRINCIPLES

GUIDING PRINCIPLES

- Making data *available* isn't enough to make it *usable.*
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- The impact of TCGA is extended by new data & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

GUIDING PRINCIPLES

- **Making data *available*** isn't enough to make it *usable*.
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- The impact of TCGA is extended by new data & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

MORE THAN ONE PETABYTE OF TCGA DATA AT YOUR FINGERTIPS

Open Data

Information NOT unique to an individual.

- **de-identified clinical data**
- **gene expression data**
- **copy number alterations**
- **epigenetic data**

Controlled Data

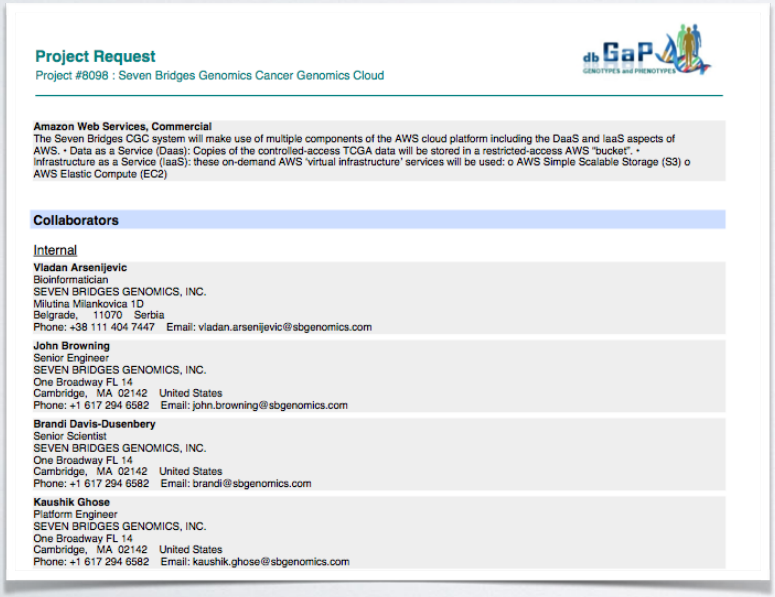
Information that IS unique to an individual.

- **primary sequencing data**
- **raw & processed SNP6 array data**
- **raw exon array data**
- **mutation calls for an individual**

CANCER GENOMICS CLOUD
SEVEN BRIDGES

ACCESSING CONTROLLED DATA

Researchers need to be authorized in dbGaP with their NIH credentials for TCGA data and are required to comply with their Data Use Certifications.



Project Request

Project #8098 : Seven Bridges Genomics Cancer Genomics Cloud

Amazon Web Services, Commercial

The Seven Bridges CGC system will make use of multiple components of the AWS cloud platform including the DaaS and IaaS aspects of AWS. • Data as a Service (Daas): Copies of the controlled-access TCGA data will be stored in a restricted-access AWS "bucket". • Infrastructure as a Service (IaaS): these on-demand AWS 'virtual infrastructure' services will be used: o AWS Simple Scalable Storage (S3) o AWS Elastic Compute (EC2)

Collaborators

Internal

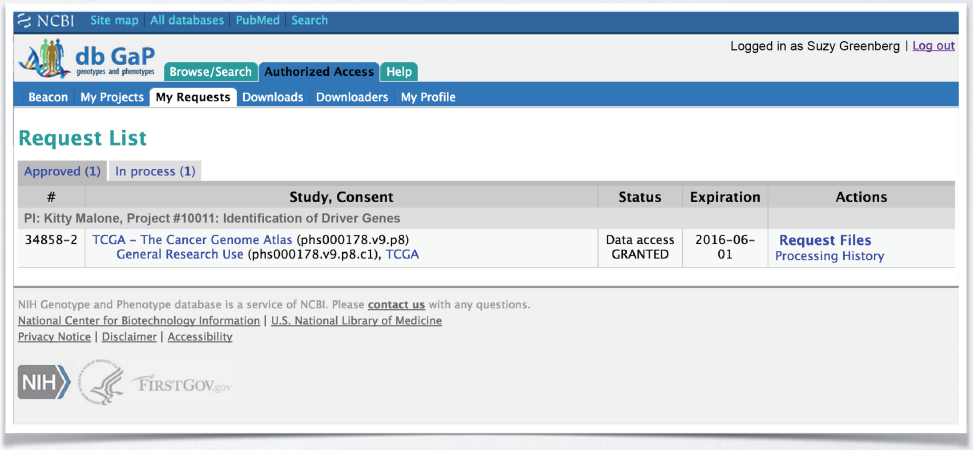
Vladan Arsenijevic
Bioinformatician
SEVEN BRIDGES GENOMICS, INC.
Milutina Milankovica 1D
Belgrade, 11070 Serbia
Phone: +38 111 404 7447 Email: vladan.arsenijevic@sbgenomics.com

John Browning
Senior Engineer
SEVEN BRIDGES GENOMICS, INC.
One Broadway FL 14
Cambridge, MA 02142 United States
Phone: +1 617 294 6582 Email: john.browning@sbgenomics.com

Brandi Davis-Dusenbery
Senior Scientist
SEVEN BRIDGES GENOMICS, INC.
One Broadway FL 14
Cambridge, MA 02142 United States
Phone: +1 617 294 6582 Email: brandi@sbgenomics.com

Kaushik Ghose
Platform Engineer
SEVEN BRIDGES GENOMICS, INC.
One Broadway FL 14
Cambridge, MA 02142 United States
Phone: +1 617 294 6582 Email: kaushik.ghose@sbgenomics.com

!=



NCBI Site map | All databases | PubMed | Search

db GaP genotypes and phenotypes

Logged in as Suzy Greenberg | Log out

Browse/Search | Authorized Access | Help

Beacon | My Projects | My Requests | Downloads | Downloaders | My Profile

Request List

Approved (1) | In process (1)

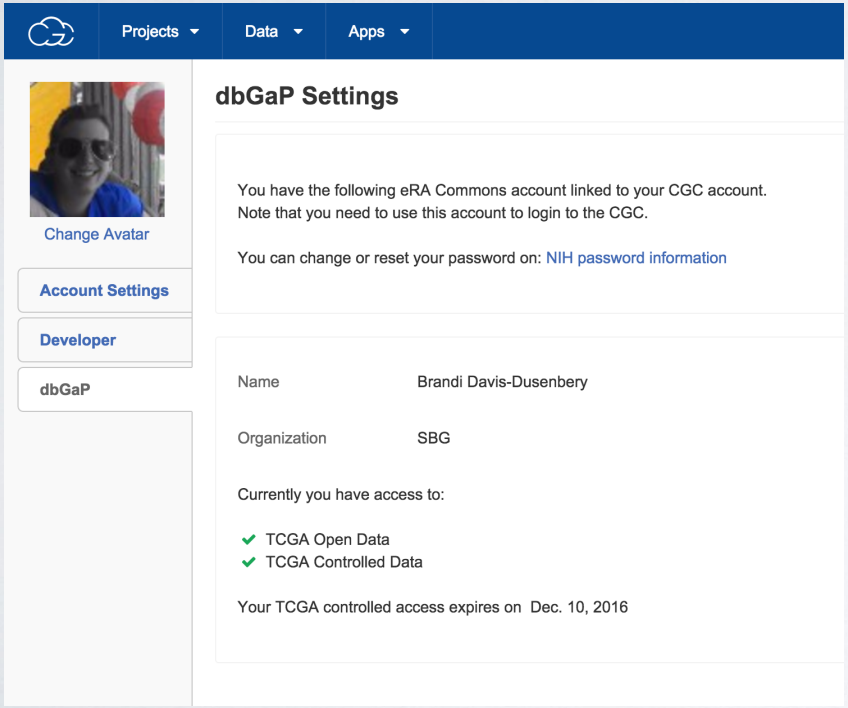
#	Study, Consent	Status	Expiration	Actions
PI: Kitty Malone, Project #10011: Identification of Driver Genes				
34858-2	TCGA – The Cancer Genome Atlas (phs000178.v9.p8) General Research Use (phs000178.v9.p8.c1), TCGA	Data access GRANTED	2016-06-01	Request Files Processing History

NIH Genotype and Phenotype database is a service of NCBI. Please contact us with any questions.
National Center for Biotechnology Information | U.S. National Library of Medicine
Privacy Notice | Disclaimer | Accessibility

NIH FIRSTGOV.gov

CANCER GENOMICS CLOUD
SEVEN BRIDGES

ACCESSING CONTROLLED DATA



- To access Controlled Data, log in with your eRA commons or NIH credentials.
- TCGA data access is verified nightly and you can always check your status.
- The email listed in eRA Commons will be used for all notifications.

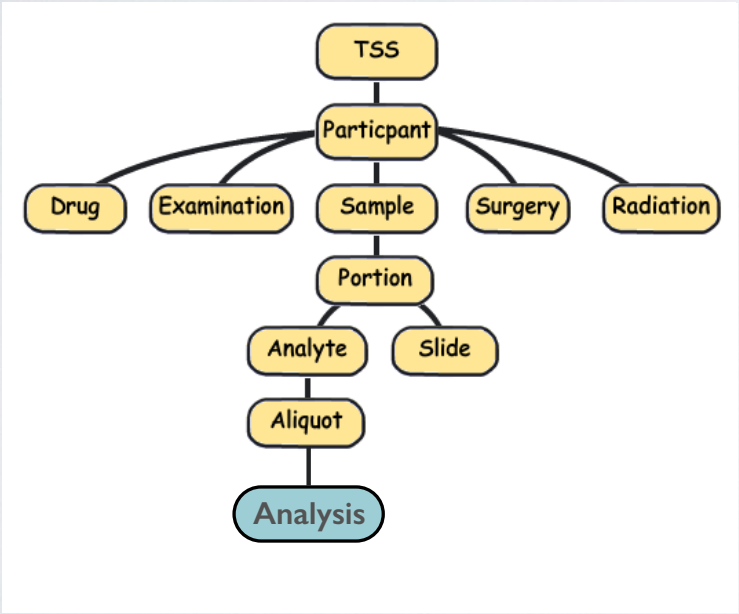
CANCER GENOMICS CLOUD
SEVEN BRIDGES

GUIDING PRINCIPLES

- Making data *available* isn't enough to **make it *usable*.**
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- The impact of TCGA is extended by new data & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

UNDERSTANDING TCGA DATASET



Biospecimen

Clinical

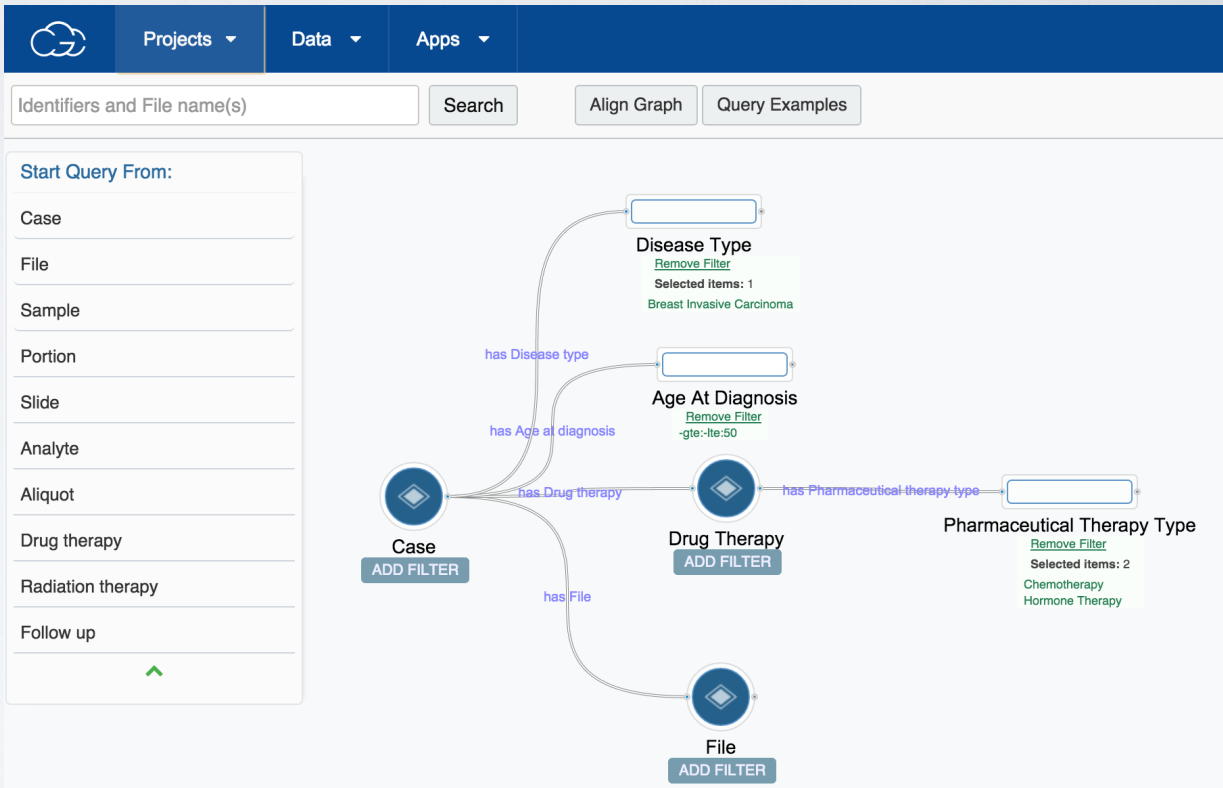
Analysis

Barcodes, and UUIDs, and XMLs, oh my

CANCER GENOMICS CLOUD
SEVEN BRIDGES

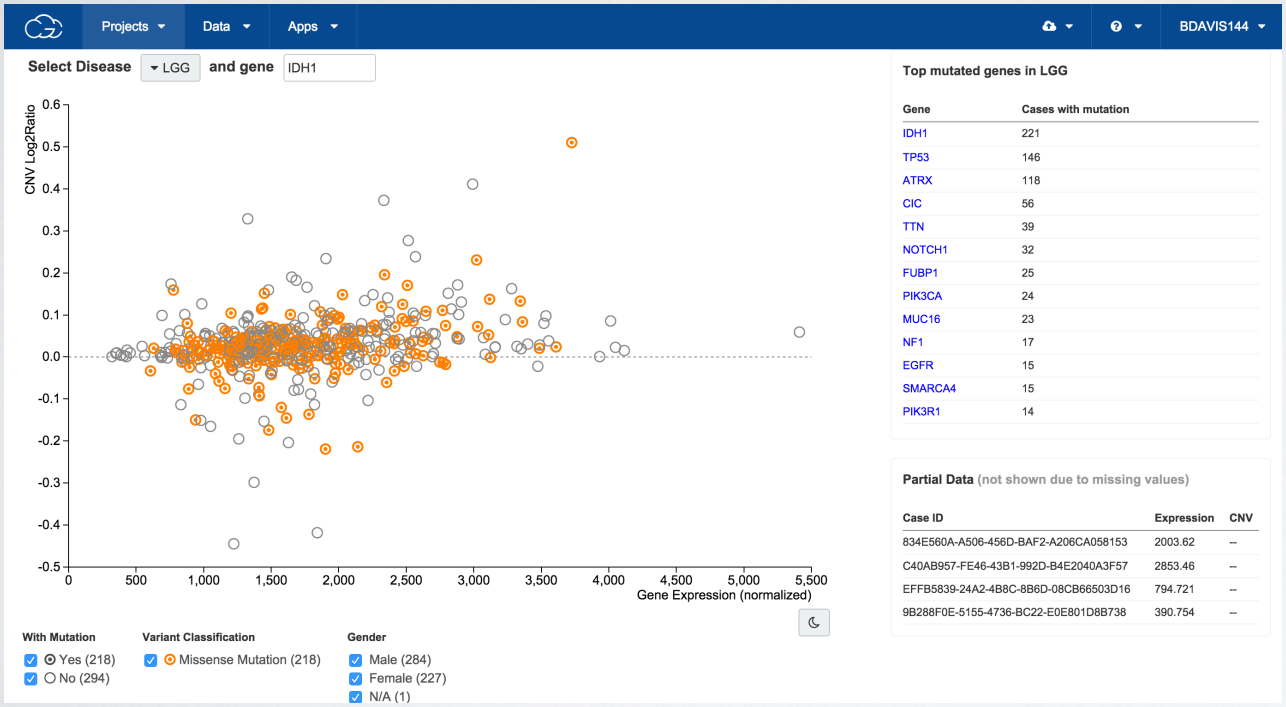
TOGETHER AT LAST

- Semantic knowledge base with >140 properties about cases, samples, files & more.
- Visual or programatic query



All files from breast cancer patients who where diagnosed under age 50 AND treated with EITHER Chemotherapy or Hormone Therapy

EXPLORE PROCESSED DATA

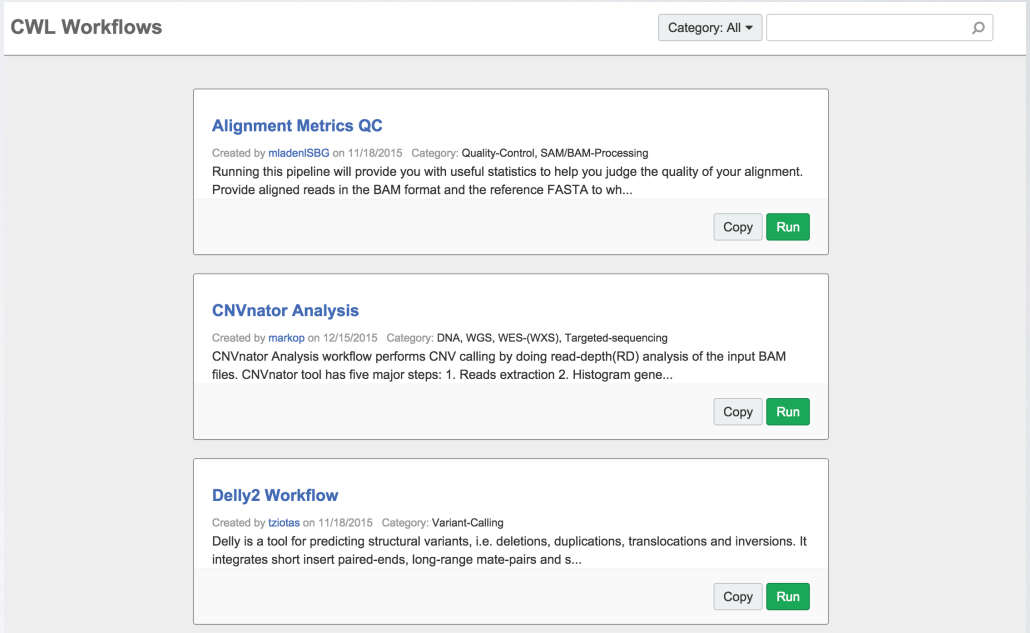


Mutations,
Copy Number Variation,
Expression Levels

CANCER GENOMICS CLOUD
SEVEN BRIDGES

IMMEDIATELY RUN AN ANALYSIS

~150 TOOLS AND WORKFLOWS ON THE CGC TODAY.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

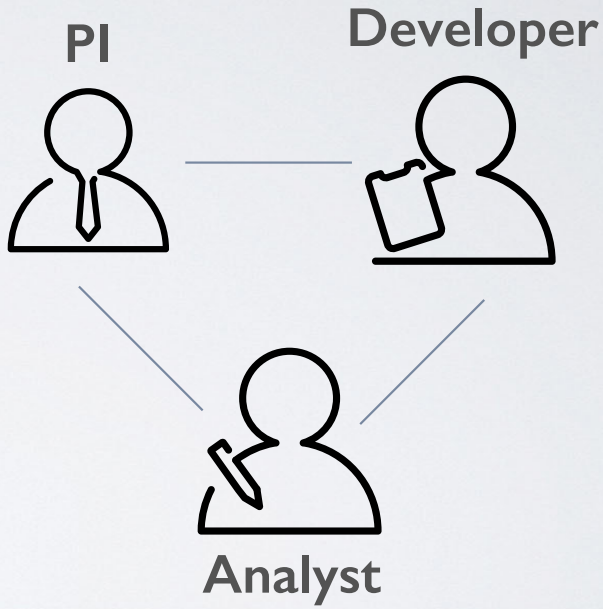
GUIDING PRINCIPLES

- Making data *available* isn't enough to make it *usable.*
- **The best science happens in teams.**
- Reproducibility shouldn't be hard.
- The impact of TCGA is extended by new data & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

EASY COLLABORATION

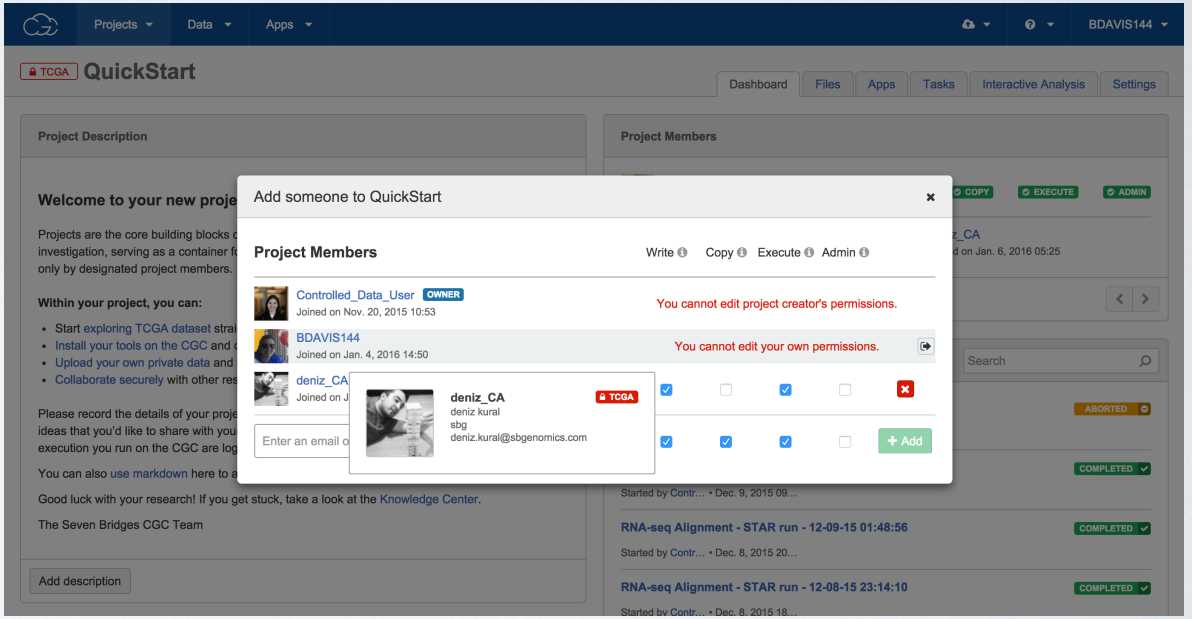
- Projects serve as shared workspaces with data and tools.
- Fine grained permissions let you set who can do what, and communicate with your team.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

COMPLIANT COLLABORATION

TCGA Controlled data projects access limited to only Authorized users.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

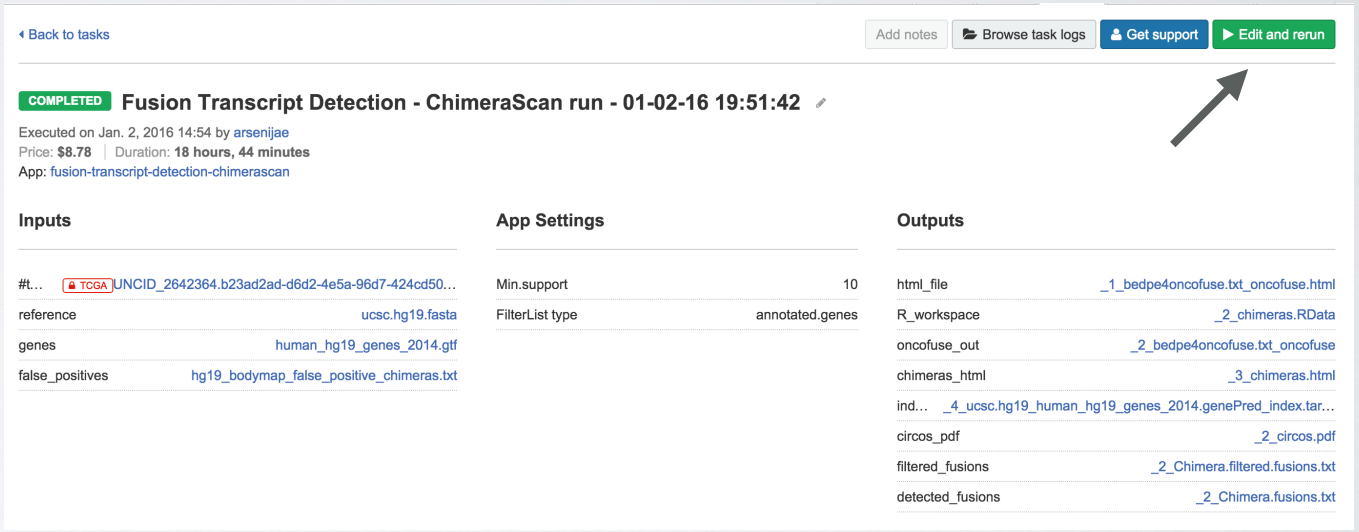
GUIDING PRINCIPLES

- Making data *available* isn't enough to make it *usable.*
- The best science happens in teams.
- **Reproducibility shouldn't be hard.**
- The impact of TCGA is extended by new data & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

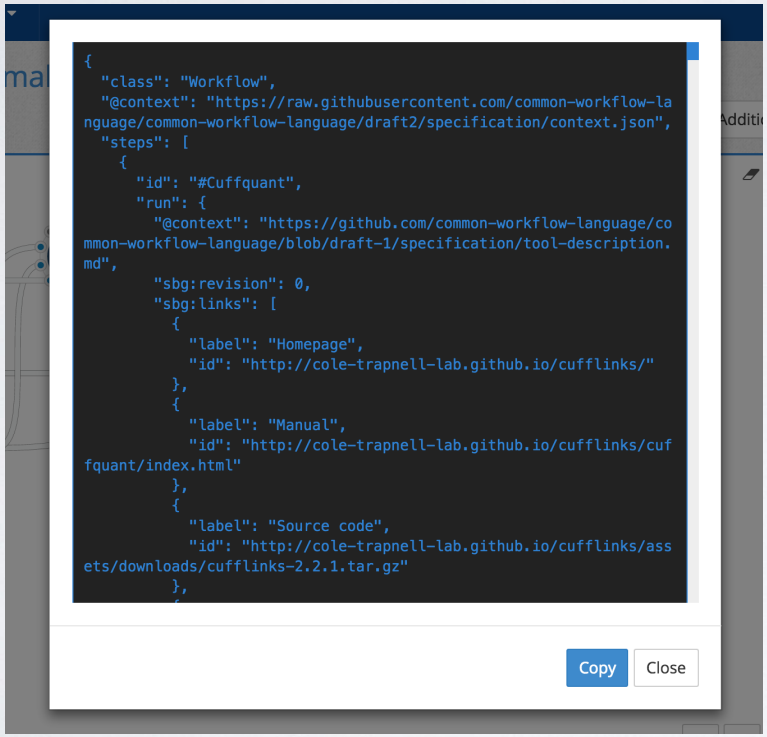
EACH TASK IS REPLICABLE & REMEMBERABLE

The inputs, outputs, and parameters as well of the precise tool versions (including dependencies!) are always linked and available for reference days or months later.



CANCER GENOMICS CLOUD
SEVEN BRIDGES

… AND SELF CONTAINED



- Even the most complex workflows are captured as small run-able text files.
- Easy to share and save.

CANCER GENOMICS CLOUD
SEVEN BRIDGES

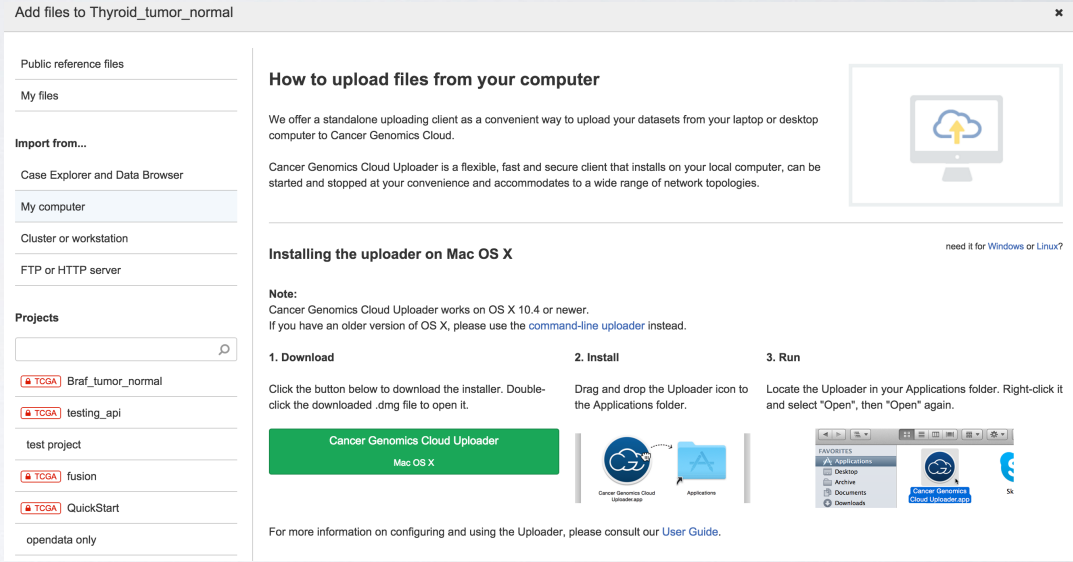
GUIDING PRINCIPLES

- Making data *available* isn't enough to make it *usable.*
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- **The impact of TCGA is extended by new data** & tools

CANCER GENOMICS CLOUD
SEVEN BRIDGES

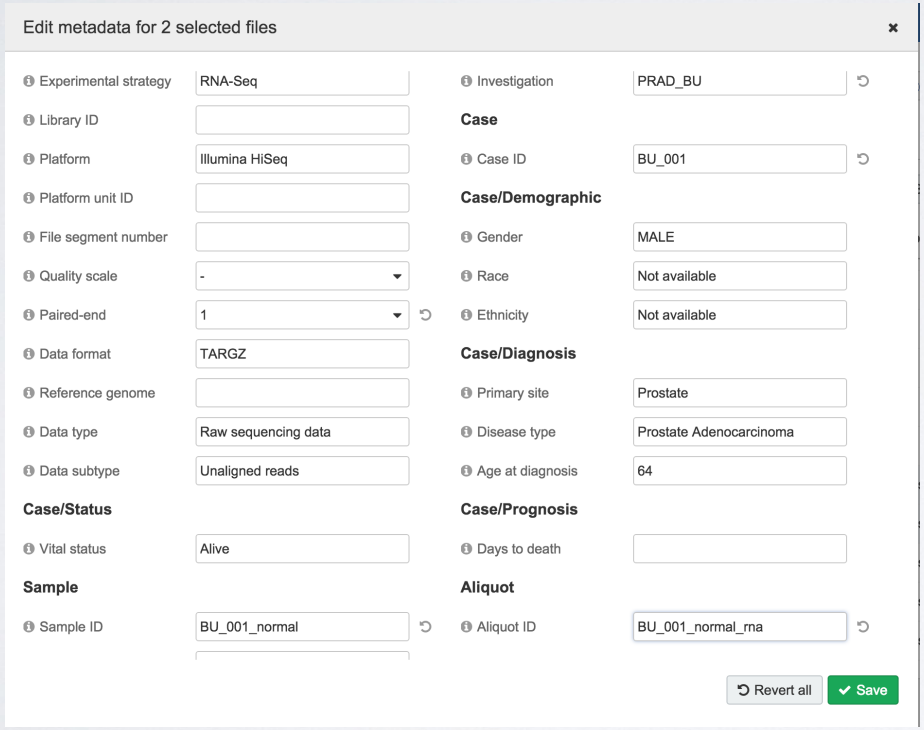
4 WAYS TO ADD DATA

- Graphical uploader
- Command Line uploader
- FTP / HTTP
- API



CANCER GENOMICS CLOUD
SEVEN BRIDGES

EASILY ANNOTATE UPLOADED DATA - SO YOU CAN FIND IT LATER



~40 properties in visual interface, unlimited custom properties via API.

CANCER GENOMICS CLOUD
SEVEN BRIDGES

GUIDING PRINCIPLES

- Making data *available* isn't enough to make it *usable*.
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- **The impact of TCGA is extended by** new data & **tools**

CANCER GENOMICS CLOUD
SEVEN BRIDGES

AS THE AMOUNT OF DATA HAS GROWN, SO TOO HAS THE NUMBER OF TOOLS AVAILABLE TO ANALYZE IT.

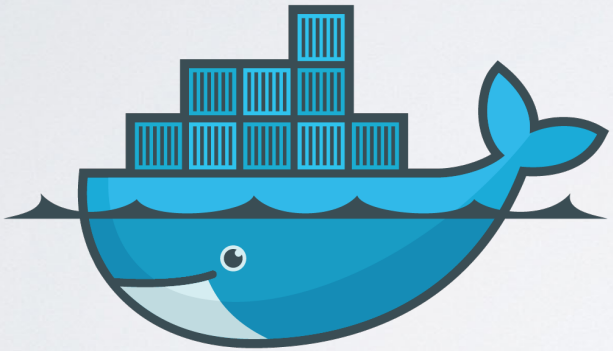
10,654 -omics data analysis tools*
(each with many versions)

50+ used in a single TCGA marker paper

*omictools.com

CANCER GENOMICS CLOUD
SEVEN BRIDGES

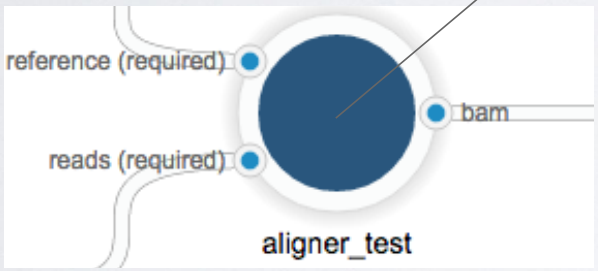
DOCKER + CWL MAKES IT EASY TO PUT THESE TOOLS ON THE CGC …AND OTHER PLACES.



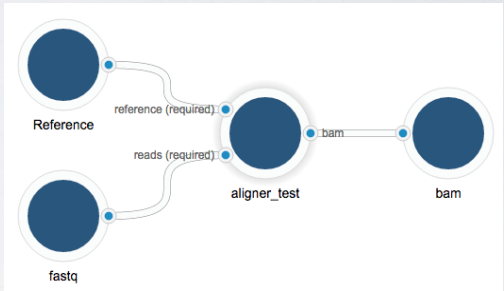
+

CANCER GENOMICS CLOUD
SEVEN BRIDGES

DEFINE THE TOOL, INPUTS OUTPUTS, AND PARAMETERS



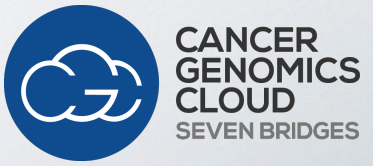

```
"softwareDescription": {
  "name": "myaligner",
  "description": "Aligns reads to a reference"
},
"documentAuthor": "kaushikghose@sbgenomics.com",
"requirements": {
  "environment": {
    "container": {
      "type": "docker",
      "uri": "",
      "imageId": ""
    }
  },
  "resources": {
    "cpu": 0,
    "mem": 5000,
    "ports": [],
    "diskSpace": 0,
    "network": false
  }
},
```



CANCER GENOMICS CLOUD
SEVEN BRIDGES

THE CGC IN ACTION

Examine gene expression differences between
Primary Tumor and Solid Tissue Normal samples
from Thyroid Cancer patients with BRAF mutation



THANK YOU!!