

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

Brandi Davis-Dusenbery, PhD
CBIIT Speaker Series
January 6, 2016

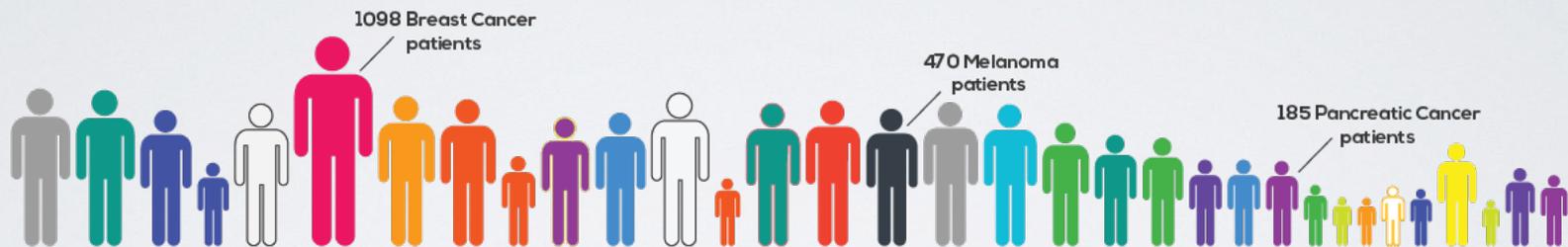


**CANCER
GENOMICS
CLOUD**
SEVEN BRIDGES

AGENDA

- Motivation
- Guiding Principles
- Case study

TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY ...



More than 11,000 cases representing 33 cancer types

TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY ...

Primary Tumor
Metastatic
...



Blood Derived Normal
Solid Tissue Normal
...

Genomic
Proteomic
...

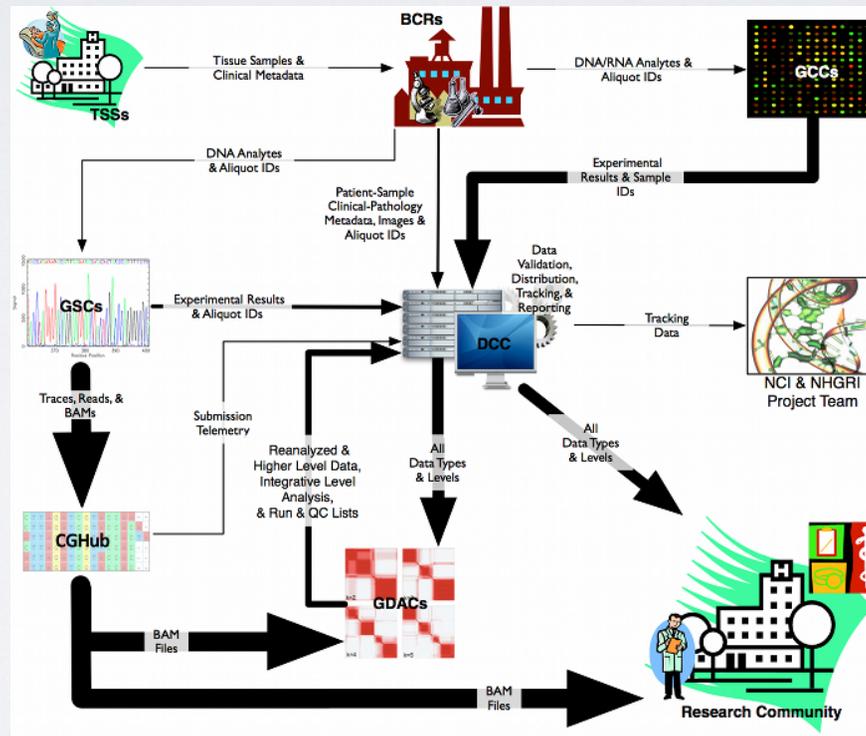


Transcriptomic
Epigenomic
...

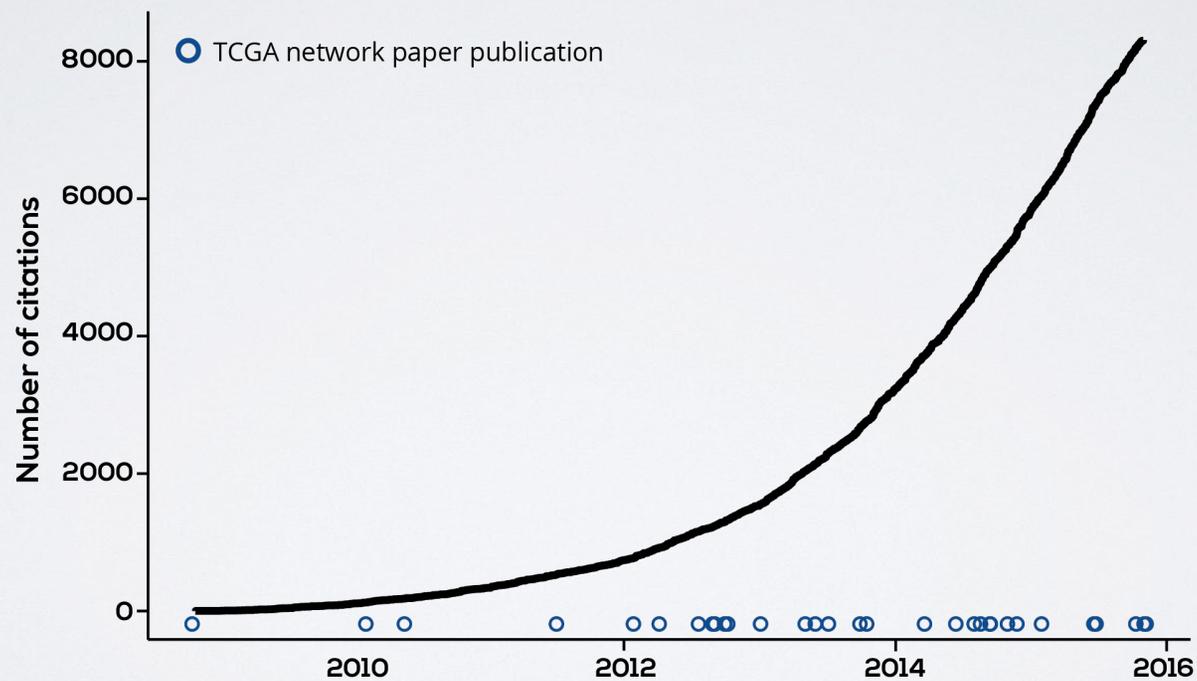
multiple Samples per Case

multiple Analyses per Sample

... MADE POSSIBLE BY
THOUSANDS OF RESEARCHERS ...



... WITH FAR REACHING
IMPACT.



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

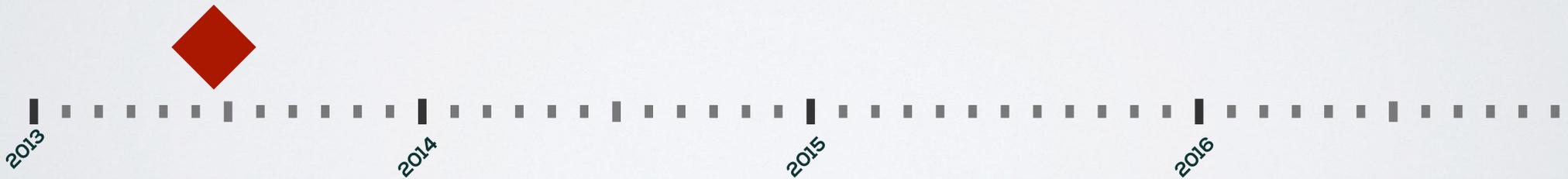
3 YEARS IN THE MAKING...

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.



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.



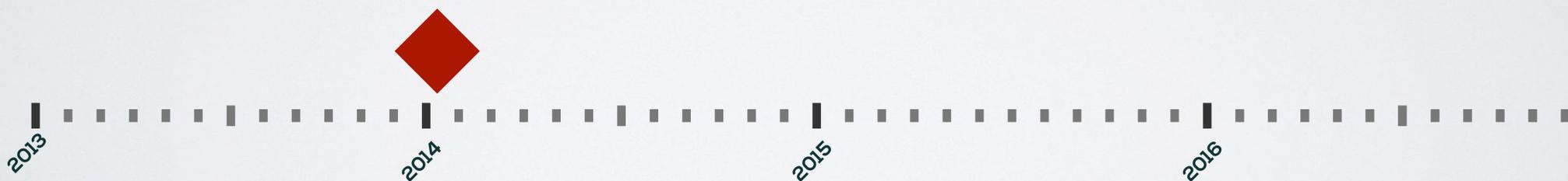
3 YEARS IN THE MAKING...

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



3 YEARS IN THE MAKING...

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



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.



GUIDING PRINCIPLES

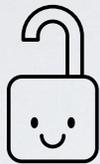
GUIDING PRINCIPLES

- 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

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

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

Viadan Arsenijevic
Bioinformatician
SEVEN BRIDGES GENOMICS, INC.
Milutina Milankovica 1D
Belgrade, 11070 Serbia
Phone: +38 111 404 7447 Email: viadan.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 Browse/Search Authorized Access Help

Logged in as Suzy Greenberg | Log out

Beacon My Projects My Requests Downloads Downloaders My Profile

Request List

Approved (1) In process (1)

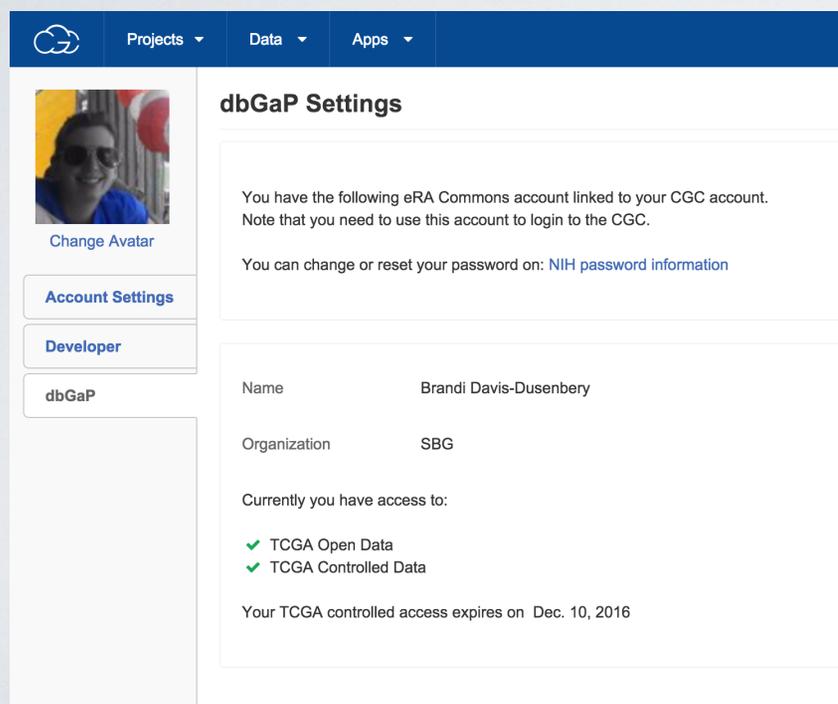
#	Study, Consent	Status	Expiration	Actions
34858-2	TCGA - The Cancer Genome Atlas (phs000178.v9.p8) General Research Use (phs000178.v9.p8.c1), TCGA	Data access GRANTED	2016-06-01	Request Files Processing History

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](#)





ACCESSING CONTROLLED DATA



The screenshot shows the 'dbGaP Settings' page. At the top, there is a navigation bar with 'Projects', 'Data', and 'Apps' dropdown menus. Below the navigation bar, there is a user profile section with a photo of a woman and a 'Change Avatar' link. To the right of the profile, the 'dbGaP Settings' section contains the following information:

- You have the following eRA Commons account linked to your CGC account. Note that you need to use this account to login to the CGC.
- You can change or reset your password on: [NIH password information](#)
- Name: Brandi Davis-Dusenbery
- Organization: SBG
- Currently you have access to:
 - ✓ TCGA Open Data
 - ✓ TCGA Controlled Data
- Your TCGA controlled access expires on Dec. 10, 2016

- 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.

GUIDING PRINCIPLES

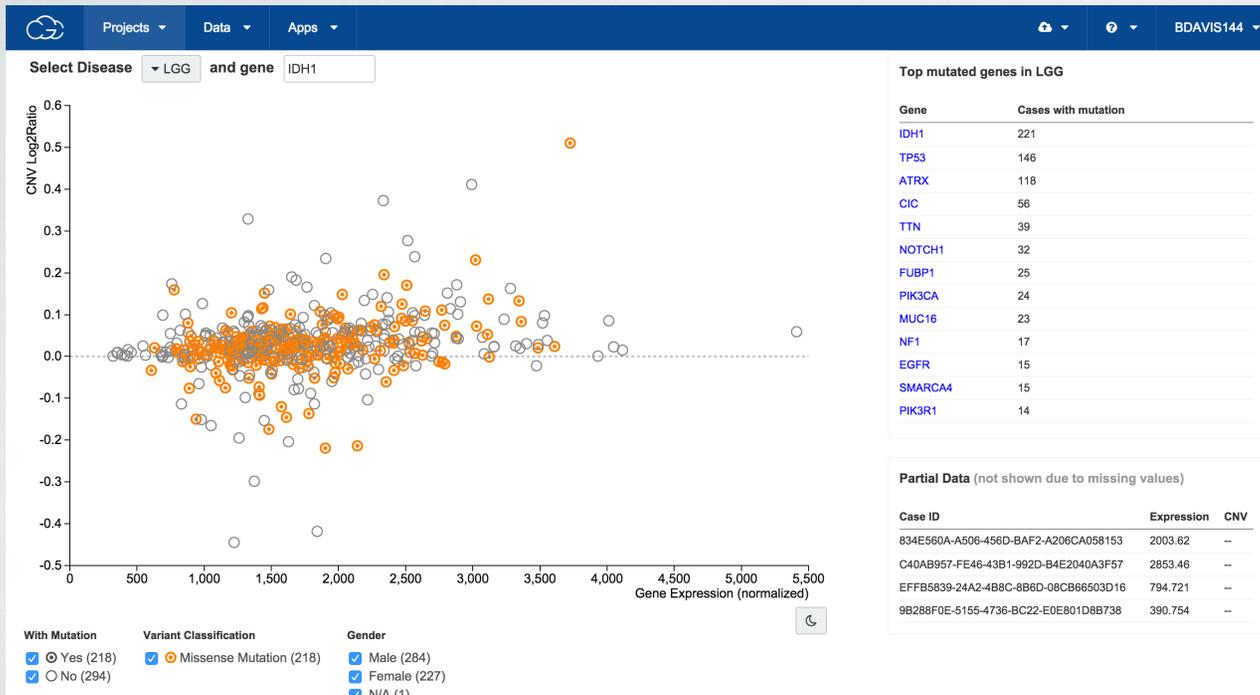
- Making data *available* isn't enough to **make it usable**.
- The best science happens in teams.
- Reproducibility shouldn't be hard.
- The impact of TCGA is extended by new data & tools

TOGETHER AT LAST

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

The screenshot displays a web-based semantic query interface. At the top, there is a navigation bar with 'Projects', 'Data', and 'Apps' menus. Below this is a search bar labeled 'Identifiers and File name(s)' and buttons for 'Search', 'Align Graph', and 'Query Examples'. The main area is divided into two panels. The left panel, titled 'Start Query From:', lists various entities: Case, File, Sample, Portion, Slide, Analyte, Aliquot, Drug therapy, Radiation therapy, and Follow up. The right panel shows a graph query visualization. The central node is 'Case', which has four outgoing relationships: 'has Disease type' leading to 'Disease Type', 'has Age at diagnosis' leading to 'Age At Diagnosis', 'has Drug therapy' leading to 'Drug Therapy', and 'has File' leading to 'File'. Each of these nodes has an 'ADD FILTER' button. The 'Disease Type' node is highlighted in green and shows 'Selected items: 1' with 'Breast Invasive Carcinoma'. The 'Age At Diagnosis' node is also highlighted in green and shows a filter '-gte:-lte:50'. The 'Drug Therapy' node is highlighted in green and has an outgoing relationship 'has Pharmaceutical therapy type' leading to 'Pharmaceutical Therapy Type', which is highlighted in green and shows 'Selected items: 2' with 'Chemotherapy' and 'Hormone Therapy'. The 'File' node is highlighted in green and has an 'ADD FILTER' button. At the bottom right, a text box summarizes the query: 'All files from breast cancer patients who were diagnosed under age 50 AND treated with EITHER Chemotherapy or Hormone Therapy'.

EXPLORE PROCESSED DATA



Mutations,
Copy Number Variation,
Expression Levels

IMMEDIATELY RUN AN ANALYSIS

~ 150 TOOLS AND WORKFLOWS ON THE CGC TODAY.

CWL Workflows Category: All

Alignment Metrics QC
Created by [mladenSBC](#) on 11/18/2015 Category: Quality-Control, SAM/BAM-Processing
Running this pipeline will provide you with useful statistics to help you judge the quality of your alignment. Provide aligned reads in the BAM format and the reference FASTA to wh...

[Copy](#) [Run](#)

CNVnator Analysis
Created by [markop](#) on 12/15/2015 Category: DNA, WGS, WES-(WXS), Targeted-sequencing
CNVnator Analysis workflow performs CNV calling by doing read-depth(RD) analysis of the input BAM files. CNVnator tool has five major steps: 1. Reads extraction 2. Histogram gene...

[Copy](#) [Run](#)

Delly2 Workflow
Created by [tziotas](#) on 11/18/2015 Category: Variant-Calling
Delly is a tool for predicting structural variants, i.e. deletions, duplications, translocations and inversions. It integrates short insert paired-ends, long-range mate-pairs and s...

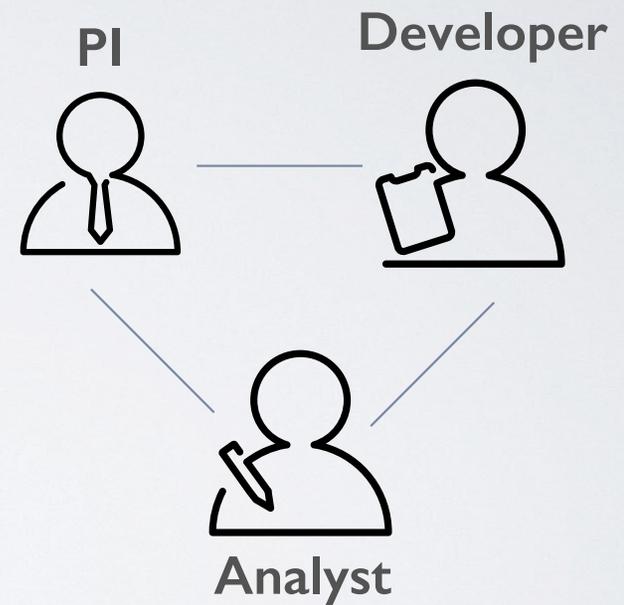
[Copy](#) [Run](#)

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

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.



COMPLIANT COLLABORATION

TCGA Controlled data projects access limited to only Authorized users.

The screenshot displays the 'QuickStart' project page in the Cancer Genomics Cloud. The page includes a navigation bar with 'Projects', 'Data', and 'Apps' menus. The main content area is divided into 'Project Description' and 'Project Members' sections. A modal window titled 'Add someone to QuickStart' is open, showing a list of project members with their names, roles, and join dates. The members listed are 'Controlled_Data_User' (OWNER), 'BDAVIS144', and 'deniz_CA'. A new member, 'deniz_CA', is being added with the email 'deniz.kural@sbgenomics.com'. The modal also shows a table of permissions for each member, with columns for 'Write', 'Copy', 'Execute', and 'Admin'. The 'deniz_CA' row has checkboxes for 'Write', 'Copy', and 'Execute' checked, and 'Admin' unchecked. A red message states 'You cannot edit project creator's permissions.' and another red message states 'You cannot edit your own permissions.'

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

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.

← Back to tasks Add notes Browse task logs Get support **Edit and rerun**

COMPLETED Fusion Transcript Detection - ChimeraScan run - 01-02-16 19:51:42 ↗

Executed on Jan. 2, 2016 14:54 by arsenijae
Price: **\$8.78** | Duration: **18 hours, 44 minutes**
App: fusion-transcript-detection-chimerascan

Inputs	App Settings	Outputs
#... TCGA UNCID_2642364.b23ad2ad-d6d2-4e5a-96d7-424cd50...	Min.support 10	html_file _1_bedpe4oncofuse.txt_oncofuse.html
reference ucsc.hg19.fasta	FilterList type annotated.genes	R_workspace _2_chimeras.RData
genes human_hg19_genes_2014.gtf		oncofuse_out _2_bedpe4oncofuse.txt_oncofuse
false_positives hg19_bodymap_false_positive_chimeras.txt		chimeras_html _3_chimeras.html
		ind... _4_ucsc.hg19_human_hg19_genes_2014.genePred_index.tar...
		circos_pdf _2_circos.pdf
		filtered_fusions _2_Chimera.filtered.fusions.txt
		detected_fusions _2_Chimera.fusions.txt

... AND SELF CONTAINED

```
{
  "class": "Workflow",
  "@context": "https://raw.githubusercontent.com/common-workflow-language/common-workflow-language/draft2/specification/context.json",
  "steps": [
    {
      "id": "#Cuffquant",
      "run": {
        "@context": "https://github.com/common-workflow-language/common-workflow-language/blob/draft-1/specification/tool-description.md",
        "sbg:revision": 0,
        "sbg:links": [
          {
            "label": "Homepage",
            "id": "http://cole-trapnell-lab.github.io/cufflinks/"
          },
          {
            "label": "Manual",
            "id": "http://cole-trapnell-lab.github.io/cufflinks/cuffquant/index.html"
          },
          {
            "label": "Source code",
            "id": "http://cole-trapnell-lab.github.io/cufflinks/assets/downloads/cufflinks-2.2.1.tar.gz"
          }
        ]
      }
    }
  ]
}
```

Copy Close

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

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**

4 WAYS TO ADD DATA

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

Add files to Thyroid_tumor_normal

Public reference files

My files

Import from...

Case Explorer and Data Browser

My computer

Cluster or workstation

FTP or HTTP server

Projects

TCGA Braf_tumor_normal

TCGA testing_api

test project

TCGA fusion

TCGA QuickStart

opendata only

How to upload files from your computer

We offer a standalone uploading client as a convenient way to upload your datasets from your laptop or desktop computer to Cancer Genomics Cloud.

Cancer Genomics Cloud Uploader is a flexible, fast and secure client that installs on your local computer, can be started and stopped at your convenience and accommodates to a wide range of network topologies.



need it for Windows or Linux?

Installing the uploader on Mac OS X

Note:
Cancer Genomics Cloud Uploader works on OS X 10.4 or newer.
If you have an older version of OS X, please use the [command-line uploader](#) instead.

- 1. Download**
Click the button below to download the installer. Double-click the downloaded .dmg file to open it.
- 2. Install**
Drag and drop the Uploader icon to the Applications folder.
- 3. Run**
Locate the Uploader in your Applications folder. Right-click it and select "Open", then "Open" again.



For more information on configuring and using the Uploader, please consult our [User Guide](#).

EASILY ANNOTATE UPLOADED DATA - SO YOU CAN FIND IT LATER

Edit metadata for 2 selected files

Experimental strategy	RNA-Seq	Investigation	PRAD_BU
Library ID		Case	
Platform	Illumina HiSeq	Case ID	BU_001
Platform unit ID		Case/Demographic	
File segment number		Gender	MALE
Quality scale	-	Race	Not available
Paired-end	1	Ethnicity	Not available
Data format	TARGZ	Case/Diagnosis	
Reference genome		Primary site	Prostate
Data type	Raw sequencing data	Disease type	Prostate Adenocarcinoma
Data subtype	Unaligned reads	Age at diagnosis	64
Case/Status		Case/Prognosis	
Vital status	Alive	Days to death	
Sample		Aliquot	
Sample ID	BU_001_normal	Aliquot ID	BU_001_normal_ma

[Revert all](#) [Save](#)

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

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**

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

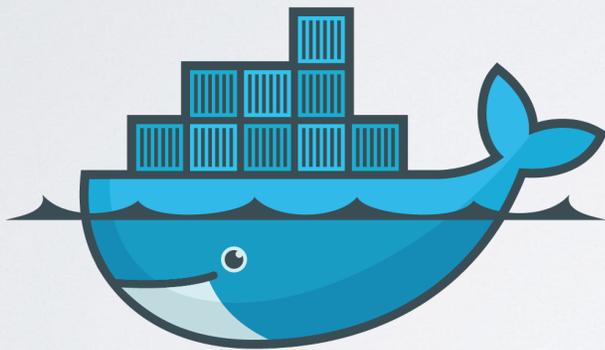
10,654 -omics data analysis tools*
(each with many versions)

50+ used in a single
TCGA marker paper

*omictools.com



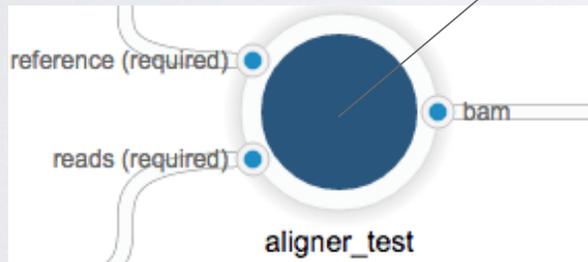
DOCKER + CWL MAKES IT EASY TO PUT
THESE TOOLS ON THE CGC ...AND
OTHER PLACES.



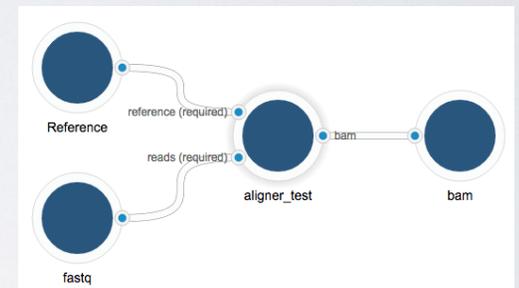
+



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  
  }  
},  
},
```



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!!