

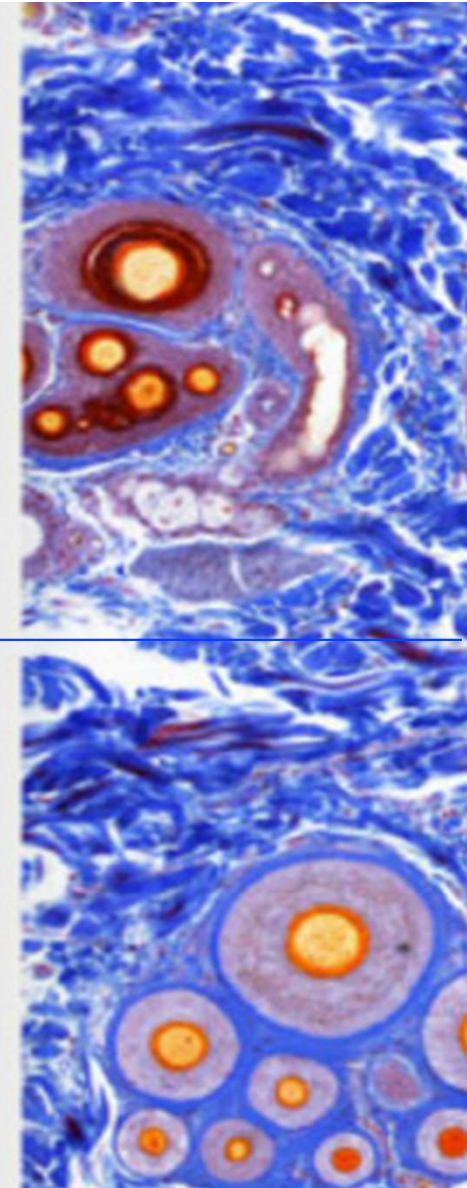
Bioinformatics in the Cloud with the CGC

Leveraging the cloud to turbocharge your research

BTEP webinar

07/21/2026

Jeni Cliffe



Mini-bio – Jenepher A. Cliffe (M.S. Astrophysics)



National Aeronautics and Space Administration
Goddard Space Flight Center



IMAGINE THE UNIVERSE!

[Home](#)
[Science](#)
[Observatories](#)
[Special Exhibits](#)
[Educators](#)
[Ask An Astrophysicist](#)
[Resources](#)
[Search](#)

What does

The best way to look at the shoes of one, but this is the next best the tools to solve many different things.

Science is a process. In these pages you are seeking.

At least one of the above methods will give you the correct answer. Which one do you want to try?

$$I = \frac{I_0}{r^2}$$

Click here to use M31's lightcurve and the $1/r^2$ relationship



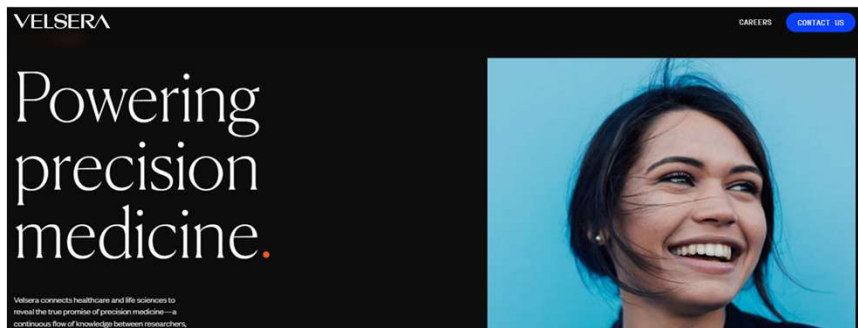
Click here to use M31's spectrum and the Doppler shift

$$H_0$$

Click here to try Hubble's Law

VELSERA

Community Engagement Manager, Velsera



- Make NIH research better by:
 - Helping scientists and doctors analyze their research data on the **Cloud**
 - Teaching people how to work with and manage **big data**
 - Giving people better **tools** so research happens faster

From exploding stars to –
humans?!

Agenda.

Bioinformatics in the Cloud with CGC

- 01** Introduction to the CGC and Cloud Computing

- 02** Accessing data and running tools (overview)

- 03** Variant analysis run-through

- 04** New feature highlights: OHIF viewer and 3D Slicer

Introduction to Cloud Computing with the Cancer Genomics Cloud

Let's take a poll!



How comfortable are you with
the cloud and cloud data?

1



Not
comfortable
*I've used it very little,
or never*

2



Somewhat
comfortable
I've used it before

3



Very comfortable
I use it all the time

Data generation is on a roll...

HC
P

Hap
Map

GTE_x

1000
Genomes

ClinGen
ClinVar

ExAC
gnomAD

UK
Biobank

All of Us
Research
Program

2001

1

Human
Genome

2010

1000

Human
Genomes

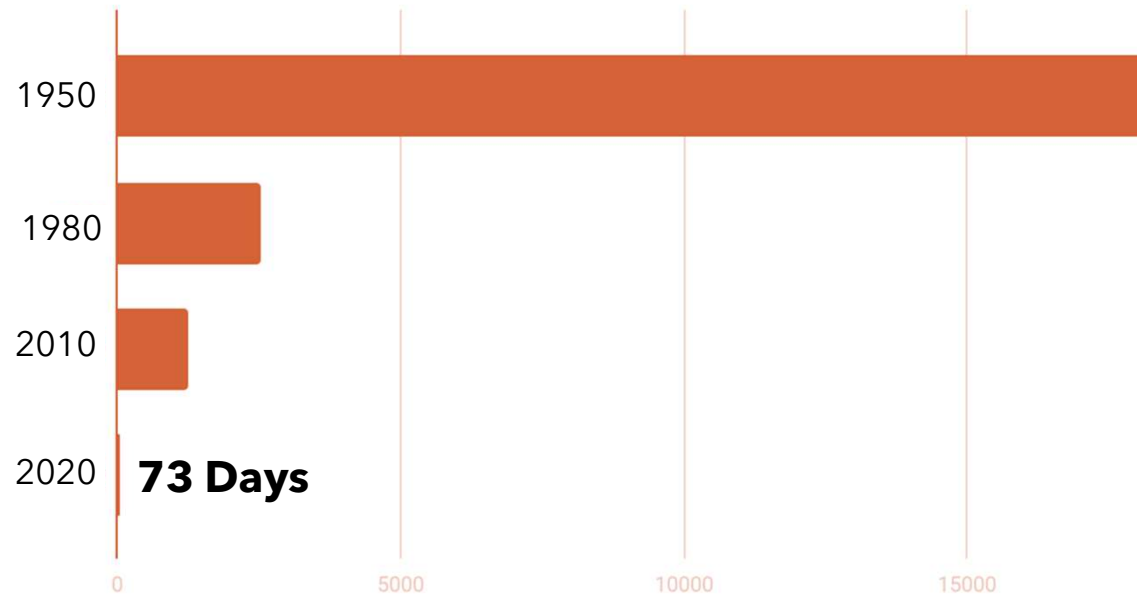
2020

1M

Human
Genomes

The Era of Big Data in Health

Doubling Time of Health Knowledge

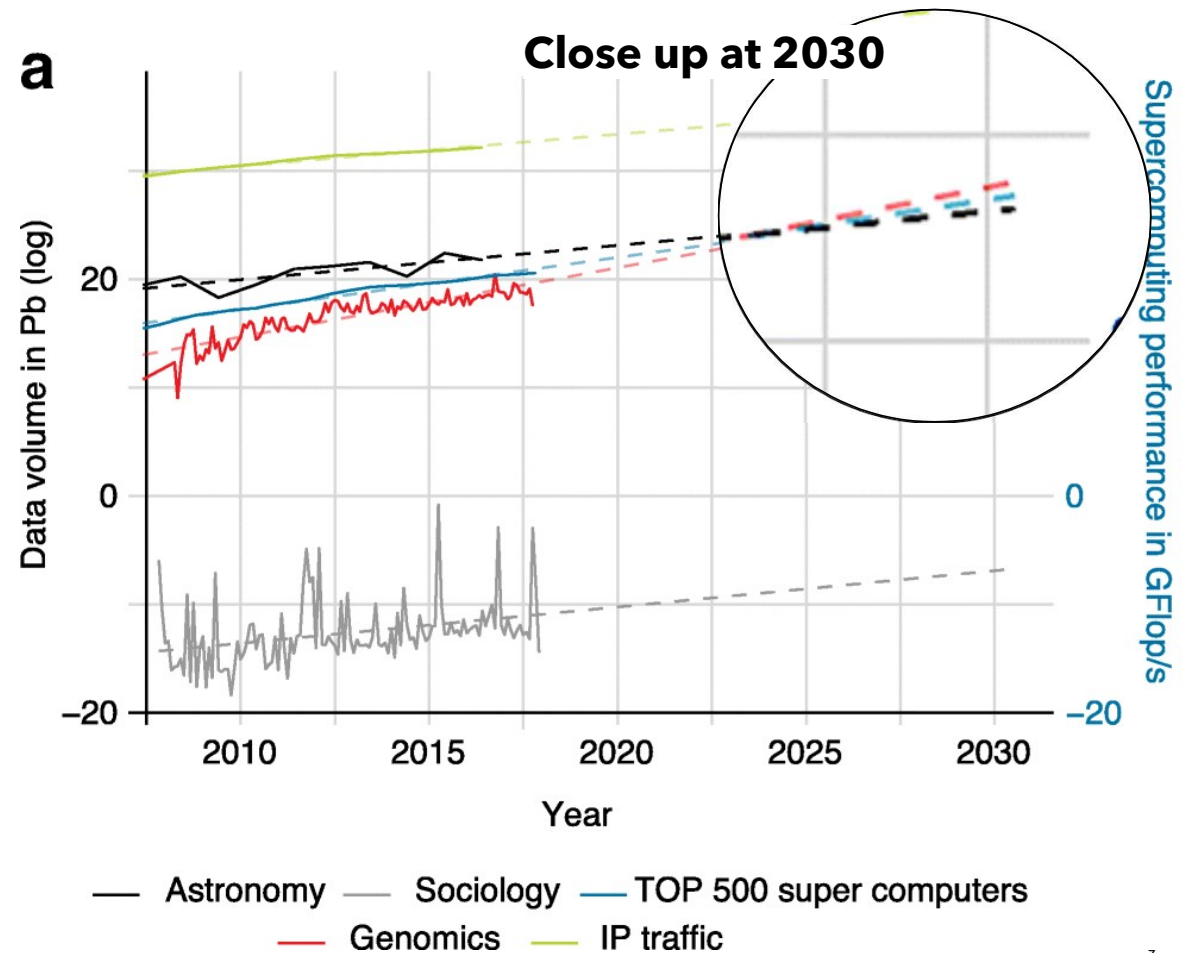


More biomedical data were generated in 2021 than all previous years **combined**

More different **kinds** of data - gene and genome sequencing, imaging, sensors (i.e. Fitbit), cellular measurement data, and more

How much data is it?

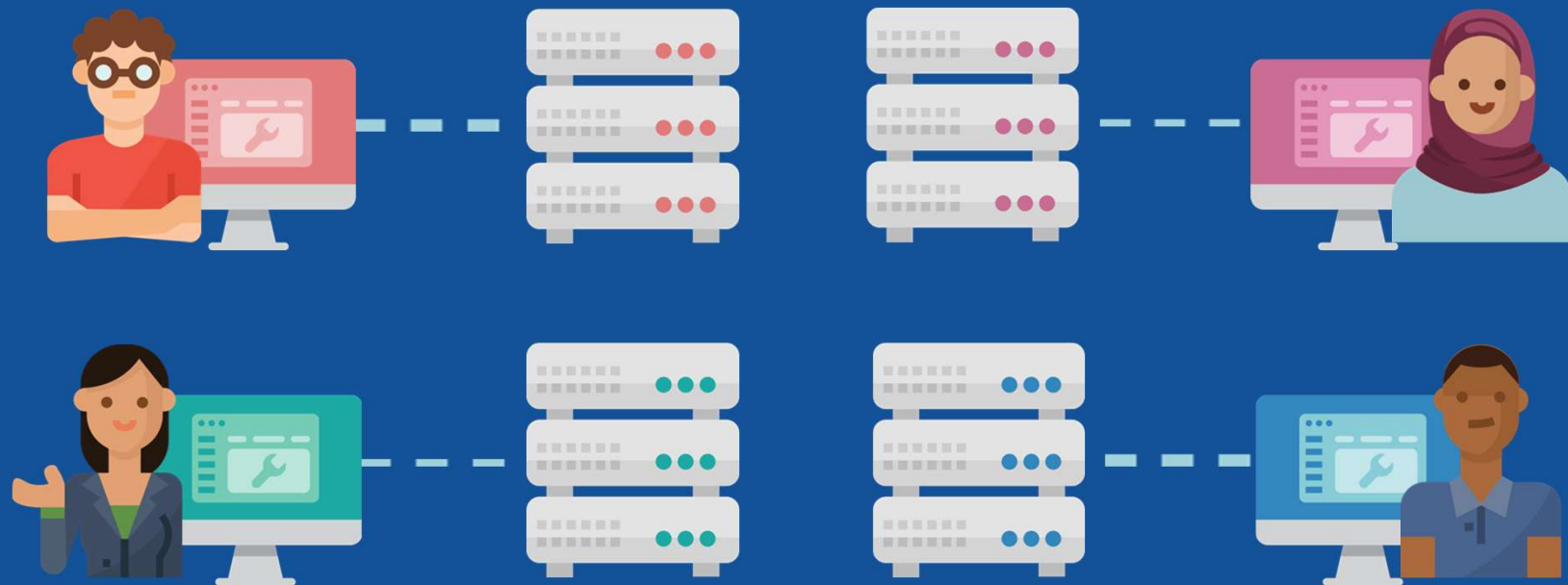
- Estimated between 2 and 40 exabytes of new data per year (Stephens *et al.* 2015)
- Genomics data is growing **faster than *the internet***
- Analysis-to-Data Paradigm - analyze data “where it lives”



How do we share and analyze datasets?

Traditional Approach

Bring data to researchers



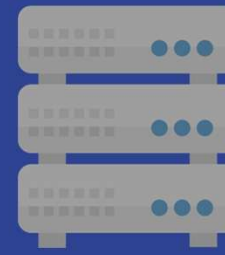
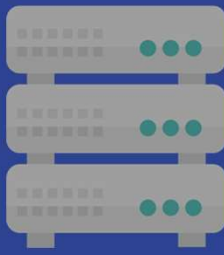
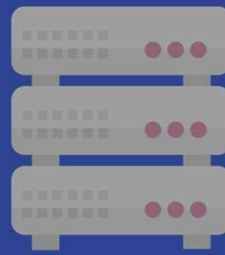
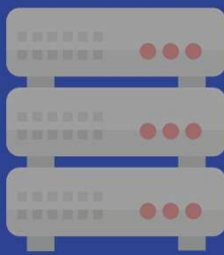
What's wrong with this traditional approach? ?

Data is
copied, not
shared

High storage &
infrastructure
costs

Hard to
share/
reproduce

Security on
individuals



Slide 11

RS1 Is this supposed to be there? Same comment on next slide, thought this was CGC.
Renee Sears, 2026-07-21T04:24:33.070

JC1 0 Nope. I hadn't gotten around to the branding stuff!
Jeni Cliffe, 2026-07-21T13:07:38.287

A new cloud-native vision for biomedical research



A new cloud-native vision for biomedical research

~~Multiple copies, not shared~~

~~High storage & infrastructure costs~~

~~Hard to share/reproduce~~

~~Security on individuals~~

True immediate data sharing

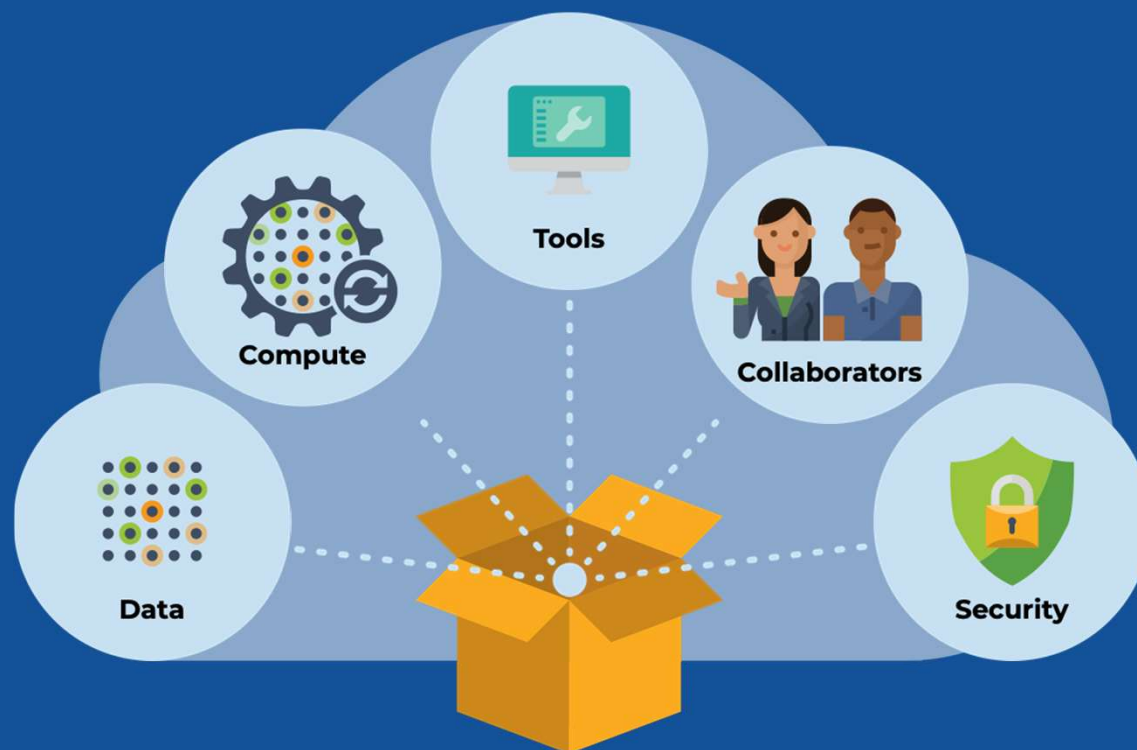
Single storage, many can access. Elastic storage & compute (pay for your use)

Shareable, reproducible tools and standardized environments

Centralized security

A new cloud-native vision for bioinformatics

**Migrating to the cloud
can be challenging!!**



Slide 14

RS1

Same comment

Renee Sears, 2026-07-21T04:24:59.180

CANCER GENOMICS CLOUD

SEVEN BRIDGES



cancergenomicscloud.org

A new cloud-native vision for biomedical research

The screenshot shows a web interface for a project titled "JAC_copy_WGS Data Analysis with GENESIS Public Project". The interface includes a navigation bar with links like Home, Projects, Data, Public Apps, Public Projects, and Developer. Below the navigation bar, there are tabs for Dashboard, Files, Apps, Tasks, Data Studio, and Interactive Apps. The main content area is divided into sections: Description, Tags, Members, and a list of completed tasks.

Access
Immediately & seamlessly (including controlled data)

Analyze
Easy to use and find curated tools and resources

Share
As easily as a Google doc, with granular permissions for control

WGS Data Analysis with G

This project is designed to introduce the user to the (SeqArray, SeqVarTools, and SNPRelate) used to perform WGS data. It consists of interactive analyses with example code that is used in GENESIS public apps, prepared with the results. Also, there are several example task apps for performing the analysis that utilize the code. Also covers use of the STAAR Pipeline public application.

The code in this project was developed as a series of exercises for the Bruce Weir Summer Institute in Statistical Genetics (SISG), and is also available on GitHub: https://uw-gac.github.io/SISG_2024.

Analyses covered in the interactive GENESIS Tutorial 2024

01_gds_intro.Rmd notebook covers:

- Converting a VCF to GDS

Members

renee_sears
Copy, Write, Execute

Completed Tasks

- COMPLETED** GDS Genotype Extractor run - 09-30-24 11:57:17
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24
- COMPLETED** STAARexercise_example_STAARpipelineSummary_VarSet_genebased
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24
- COMPLETED** STAARexercise_example_STAARpipelineSummary_IndVar_gene_centric
Submitted by: jenepher_cliffe_era · Jul 20, 2026 12:24

VELSERA



CANCER GENOMICS CLOUD

SEVEN BRIDGES

Powerful, yet easy to use interfaces to empower cancer researchers to draw new insights from petabyte scale data.

Stable, secure, and highly customizable cloud storage and computing platform

3+

Petabytes
Public Data

2.4M+

Years of
Compute

1300+

Public Tools
& Workflows

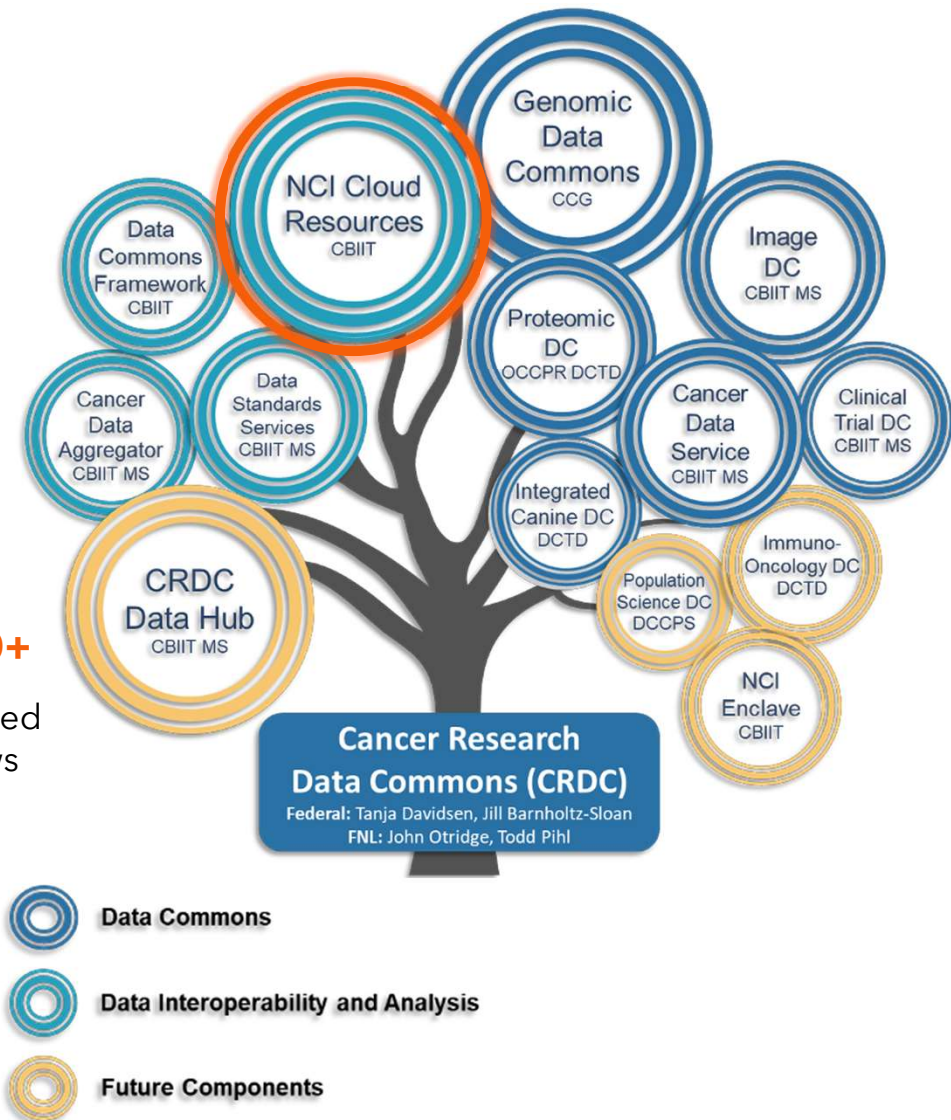
12,500+

Users

109,000+

User-Created
Workflows

VELSERA



Who are CGC Users?

The CGC is designed to serve a wide range of scientists and users with varying skill sets



BIOINFORMATICIANS

- Store, manage, and share data
- Access public and proprietary datasets
- Query, build, and investigate cohorts of interest
- Access optimized tools and workflows
- Create, optimize, maintain, and distribute new tools and workflows
- Create push-button automation solutions
- Analyze data at scale with tools and workflows
- Conduct interactive exploratory analyses
- Explore/visualize results/gather insights
- Easily collaborate with stakeholders
- Integrate with external systems



BENCH SCIENTISTS

- Store, manage, and share data
- Run optimized tools/workflows at scale
- Conduct defined analyses via push-button solutions
- Investigate/visualize results
- Easily collaborate with other stakeholders



ADMINISTRATORS

- Manage and control users
- Monitor and control institutional assets
- Manage and monitor projects
- Monitor and control costs
- Create reports



CLINICIANS

- Conduct validated analyses via push-button solutions
- Query, build, and investigate cohorts of interest
- Create reports
- Investigate/visualize results
- Easily collaborate with other stakeholders



DEVELOPERS

- Create, optimize, and maintain new tools and workflows
- Create push-button automation solutions
- Create custom interfaces for specific use cases
- Distribute proprietary tools/workflows
- Integrate upstream/downstream

RS1

Slide 18

RS1 This is a word slide, might not have enough time to bring this down now but we should think about how to restructure this for the future.

Renee Sears, 2026-07-21T04:26:47.399

JC1 0 If I have time I am going to send all the bits to Claude or Gemini to pare down. I may not have time.

Jeni Cliffe, 2026-07-21T13:12:08.810

Navigating the Cancer Genomics Cloud

“Projects” are the main unit of Organization

The screenshot shows the AIDR - Project Template interface. The top navigation bar includes links for Projects, Data, Public Apps, Public Projects, and Developer. The main content area is divided into a left sidebar with 'Description' and 'Analysis' tabs, and a right sidebar with 'Interactive Browsers', 'Settings', and 'Notes'. The 'Description' tab is active, showing 'Guidelines for Challenge Submission' and a list of tasks. The 'Analysis' tab is also visible, showing a search bar and a message about executions. A callout box in the center lists items that projects can contain.

Projects can contain

- Files
- Apps and workflows
- Collaborators
- Interactive sessions
- Notes

Description

AIDR submission

Guidelines for Challenge Submission

These are taken from the "How to Enter" tab at the Challenge.gov challenge: [Challenge.Gov](#)

A. Analyze Data and assess its AI Readiness

- Qualitative assessment (through a manual review of documentation and governance metrics that outputs a description of the current landscape)
- Quantitative assessment (through a script that outputs some metrics measures such as data completeness and consistency)

B. Process the data to improve its readiness and build an AI/ML model

- Resolve any data issues and present data in a higher AI/ML readiness as needed (through a "ml_model" script)
- Build an AI/ML model that takes the processed data and produce a model (through the same "ml_model" script)

Analysis

Search

Tasks Data Studio

Your executions will appear here. Before you start, [learn more about them.](#)

Access Public Data, or Use Your Own

 Projects ▾ Data ▾ Public Apps ▾ Public Projects Developer ▾  rowan_beck_era ▾

Dashboard **Files** Apps Tasks Data Studio **Demonstration: Building an App**  Interactive Browsers Settings Notes

 Files

 New folder  Add files ▾ 

Extension: All ▾ Sample ID: All ▾ Task ID: All ▾ Tags: All ▾  Clear filters

☐ ▾ ▲ Name	Task ID	Created on	Extension	Size
☐ G20479.HCC1143.2_1Mreads.tar.gz TEST	-	Aug. 10, 2023 11:44	TAR.GZ	115.0 MiB
☐ G20479.HCC1143.2_1Mreads_pe_1.fastq	8f6e866e-b767-45af-...	Aug. 10, 2023 11:46	FASTQ	232.2 MiB
☐ G20479.HCC1143.2_1Mreads_pe_2.fastq	8f6e866e-b767-45af-...	Aug. 10, 2023 11:46	FASTQ	232.2 MiB



Case Explorer and Data Browser

Public Files

Projects 

Your Computer

FTP / HTTP

GA4GH Data Repository Service (DRS)

Data Tools

Volumes

Import from a manifest file

Slide 21

RS1

What is this screenshot telling me with the + Add files? Is the + add files the important part? If so I would maybe animate a blown up view of that.

Renee Sears, 2026-07-21T04:37:48.183

Files and file management

The screenshot shows the Velsera web interface. The top navigation bar includes 'Home', 'Projects', 'Data', 'Public Apps', and 'Public Projects'. The 'Files' tab is selected in the secondary navigation bar. On the left, a sidebar shows the file hierarchy: 'Root / input_data' and 'Reference_files'. The main area displays a list of files under the 'input_data' folder. A callout box highlights the features of the Files tab.

The Files tab has

- Folders
- Text search
- Filter by metadata
- Filter by Tags

Files

- input_data
- Reference_files

Name
C835.HCC1143.2.converted.realigned.base_recalibrated.bam.bai
C835.HCC1143.2.converted.realigned.base_recalibrated.bam

LEARN

File Management:
<https://youtu.be/Z4ygBEBunuY>
Knowledge Center:
[Files and metadata
\(cancer.genomicscloud.org\)](https://cancer.genomicscloud.org/files-and-metadata)

No coding required to run an analysis

The screenshot displays the Galaxy web interface for a task titled "Differential Expression - Salmon + DESeq2 run - 11-30-23 17:12:19". The interface is organized into three main sections: Inputs, App Settings, and Output Settings.

Inputs:

- Batching:** Off (toggle switch).
- FASTQ read files:** A list of files including SRR9058997_1.fastq, SRR9058993_1.fastq, SRR9058992_2.fastq, SRR9058992_1.fastq, and SRR9058991_2.fastq, with a "Change selection" button.
- GTF annotation:** GRCh38ERCC.ensembl95.gtf, with a "Change selection" button.
- Genome FASTA:** No files selected, with a "Select file(s)" button.

App Settings:

- DESeq2 (#deseq2_1_26_0):** A dropdown menu showing "Genotype" as the selected covariate of interest.
- Factor level - reference:** A dropdown menu showing "WT" as the selected reference factor level.
- Factor level - test:** A dropdown menu showing "KD" as the selected test factor level.

Output Settings:

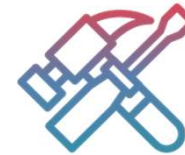
Output Setting	Value
DESeq2 analysis results	No value
Expression matrix genes	No value
Expression matrix transcripts	No value
Gene-level quantification	No value
HTML report	No value
HTML reports	No value
Normalized counts	No value
RData file	No value
Report zip	No value
Salmon Quant archive	No value
Salmon quant log	No value
Transcript-level quantification	No value
pheno out	No value

The interface also includes a "Run" button and an "Activity Monitor" button at the bottom.

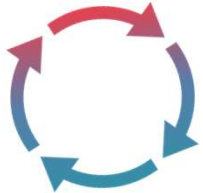
Over 1300 ^{RS1} curated tools and workflows



High quality apps & documentation



Customized user tools/workflows



Updated regularly



Secure



Optimized to run on cloud

Variant Filtration
SAM/BAM-Processing
Copy Number Analysis
Quantification
Structural Variant Calling
Quality Control
Proteomics
Gene Fusion
FASTA Processing
BED Processing
FASTQ Processing
Long Reads
RNA-Seq
Imaging
ChIP-Seq
Differential Splicing
Variant-Calling
Immune Epitopes
Indexing
Alignment
Annotation
Single Cell
Metagenomics
File Format
VCF Processing
sam/BAM Processing
RNA-Seq
ChIP-Seq
Differential Splicing
Variant-Calling
Immune Epitopes
Indexing
Alignment
Annotation
Single Cell
Metagenomics
File Format

Apps and Workflows:

<https://youtu.be/ayCKOafCGB0>

Knowledge Center:

[Workflows and tools
\(cancergenomicscloud.org\)](https://cancergenomicscloud.org)

Slide 24

RS1 Where is this number from? I see 1070 on CGC.
Renee Sears, 2026-07-21T04:28:05.118

JC1 0 I asked Bix for updated numbers
Jeni Cliffe, 2026-07-21T13:09:54.430

RS1 1 And they said 1300?
Renee Sears, 2026-07-21T17:34:15.846

Data Studio – Interactive Data Analysis

Data Studio Workbench

RS1

Derive new insights using interactive analysis environments with Data Studio environments. Code in Python, SAS, and R to record and share your analyses.

Interact with files directly using our OHIF image viewer and Galaxy.

Videos:

R Studio on the CGC: https://youtu.be/_f8DWRTS0rE

JupyterLab on the CGC: <https://youtu.be/iycWVHhggXg>

Documentation:

About Data Studio (cancergenomicscloud.org) | <https://docs.cancergenomicscloud.org/docs/about-data-cruncher>

Slide 25

RS1 Is this old terminology for Data Studios or still in use?
Renee Sears, 2026-07-21T04:35:35.495

Apps – Bulk Processing Analysis

Command Tools and Workflows

RS1

- Convert any R or Python or bash script into an app that can be optimized and automated.
- Automated and optimize to scale up.



Videos:

CGC Onboarding:

<https://youtube.com/playlist?>

Public Apps Gallery:

<https://youtu.be/ayCKOafCGB0?si=ps0O3ihGvgNvanB>

Documentation:

Apps on the CGC

(cancergenomicscloud.org) |

<https://docs.cancergenomicscloud.org/docs/apps-on-the-cgc>

Slide 26

RS1 Is this old terminology for Data Studios or still in use?
Renee Sears, 2026-07-21T04:35:35.495

Democratizes Complex Analyses



Easy data management



Secure collaboration & managed billing



Flexible & fully reproducible methods



Optimized bioinformatics algorithms



Scalable computation



Extensible & developer friendly tools

A FAIR data ecosystem

- **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR)
- Immediate scaling (“Rent” vs “Buy”)
- Levels the playing field across institutions
- Many researchers can access data - no duplicate copies.
- Reproducibility - data and methods together.

Slide 27

RS1 What does? The CGC? NIH? What is this referring to?
Renee Sears, 2026-07-21T04:36:27.010

JC1 0 CGC and cloud platforms generally...
Jeni Cliffe, 2026-07-21T13:14:53.242



HANDS-ON DEMO

cgc.SBGenomics.com

Variant Calling

Coming soon to the Cancer Genomics Cloud

Quantitative QC for Lung CT: 3D slicer

Dashboard Files Apps Tasks Data Studio Interactive Apps Copy of COVID-19 Image Segmentation with Deep... Interactive Browsers Notes

Description

This project provides deep learning image segmentation notebook, along with a pretrained model for segmentation of lung CT images. The model was trained on a publicly available dataset. Analysis files include:

- A Jupyter notebook walking through the model. This notebook can be repurposed for other applications by replacing the image preprocessing code and datasets.
- A Keras model based off of a UNet architecture with an encoder pretrained with ImageNet weights.
- Images and masks in NIfTI (.nii) format, including masks for ground glass opacities and consolidation.
- Images and lung masks in NumPy array format. The images are shuffled from the order in the dataset.
- A plot of loss history over the course of training. This will be overwritten if a new model is trained.
- Utility and dependency files.

To use the tools, select **Interactive Analysis > Data Cruncher > JupyterLab**. The notebook can be run using CPU, single-GPU, or multi-GPU instances in the **SB Machine Learning environment**. To use the model as is for predictions,

What it does

- Computes spatial overlap and boundary metrics
- Quantifies volume bias and over-segmentation
- Categorizes masks into pass, review, or fail
- Generates auditable segmentation quality records

Email notifications

yachna
Copy, Write, Execute

yachnajain
Copy, Write, Execute

jenepher_cliffe_era_login
Copy, Write, Execute



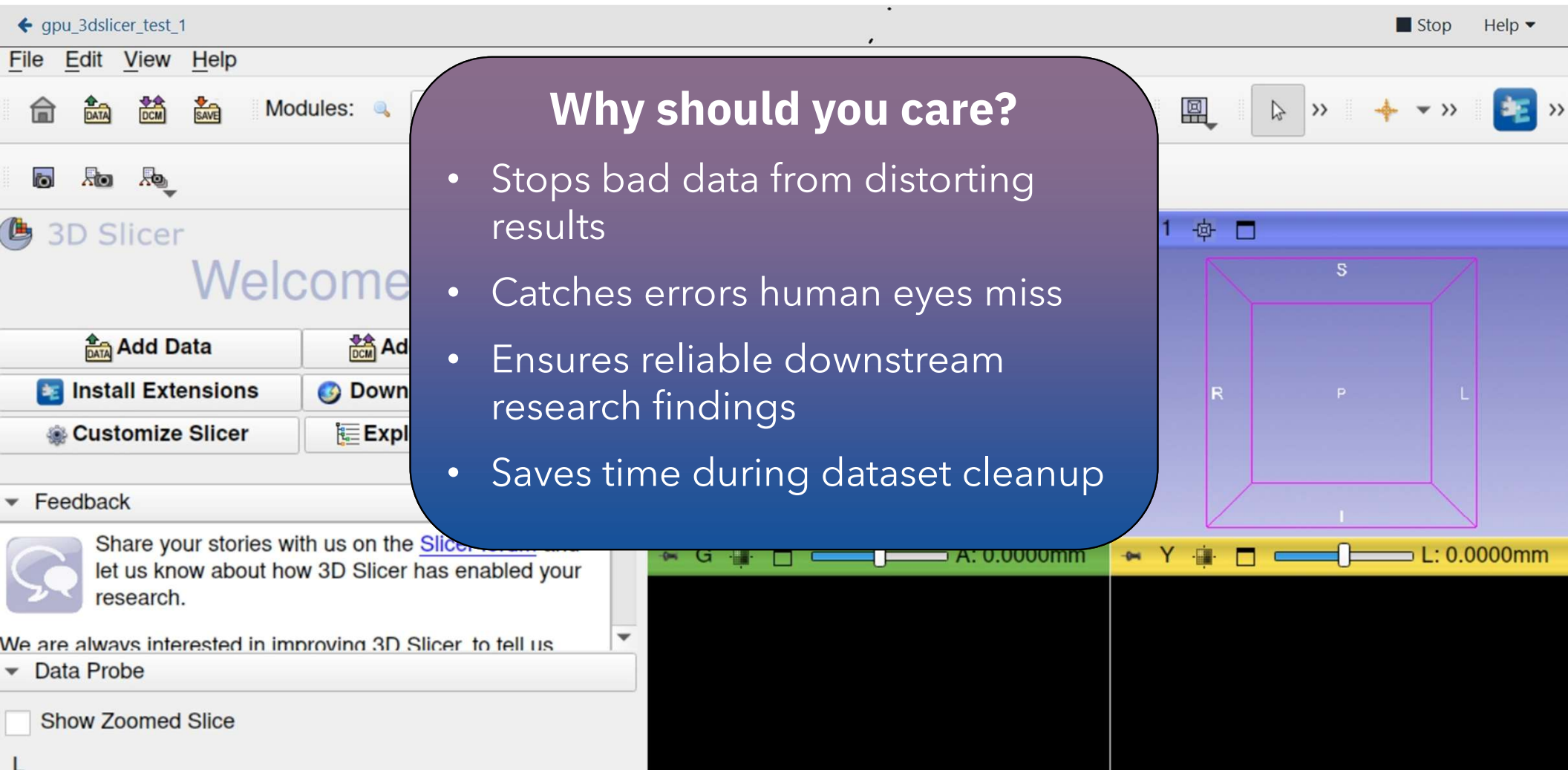
Search

Tasks Data Studio

COMPLETED SBG Compressor CWL1.0 run - 06-29-26 07:48:48

Submitted by: yatishp · Jun 29, 2026 3:48

Quantitative QC for Lung CT: 3D slicer



OHIF viewer: Data-Driven QC for Brain MRI

Dashboard Files Apps Tasks Data Studio Interactive Apps the-cancer-imaging-archive-tcia Interactive Browsers Notes

Description

The Cancer Imaging and Public Project

This project contains radiological images from The Cancer Imaging Archive (TCIA), focusing on collection Atlas (TCGA). This data is made available for research and is focused on connecting genotypes to phenotypes. The dataset includes radiological images and associated metadata stored in the standard DICOM format. Files are identified to ensure that they are freely available and they are available for public use.

Use the data in this project

The TCIA data files in this project are ready to be explored. You can explore them by using the file filters to select files based on their specific extensions or by modality using the "Modality" metadata field. For example, if you are interested in MRI images from patients with Glioblastoma Multiforme, you can filter for files with "MR" in the "Modality" metadata field.

What it does

- Standardizes MRI quality control workflows
- Validates DICOM metadata against physical measurements
- Measures tumor dimensions and volume
- Quantifies image noise and artifacts
- Exports machine-readable DICOM Structured Reports

Email notifications

yachnajain
Copy, Write, Execute

yatishp
Copy, Write, Execute

jenepher_cliffe_era_login
Copy, Write, Execute



Search

COMPLETED

FSL Brain Segmentation run - 07-01-26 10:03:39

Submitted by: anushuab · Jul 1, 2026 7:52

OHIF viewer: Data-Driven QC for Brain MRI

Dashboard

Files

Apps

Tasks

Data Studio

Interactive Apps

the-cancer-imaging-archive-tcia

Interactive Browsers

Notes

Search names and description

Create app

+ Add apps

Name

FSL Brain Segmentation

This workflow is used to analyze brain M

SBG TCIA Archive Viewer

The TCIA Archive Viewer consists of two

Cost Estimator: All

Modifi...

huab

Jul 01, 202...

Run

huab

Jul 01, 202...

Run

Showing 1 - 2 of 2

Why should you care?

- Prevents silent model training failures
- Ensures reproducible segmentation dataset quality
- Eliminates inconsistent multi-site image data
- Automates downstream quality audit trails
- Saves manual dataset review time



INTEGRATED TOOLS ENABLING

- Streamlined data access
- Reproducible
- High-throughput gating for AI pipelines
(garbage in, garbage out protection)
- Democratized computational oncology

Resources for Working on the Cancer Genomics Cloud

You are not alone!

Onboarding resources

- **\$300 Getting started videos on YouTube** RS2
- Extensive Knowledge base



Public apps and projects

- Curated tutorials and templates to adapt to your data.
- Vetted and tested apps and tools

New Public Projects (Template projects)	37
Public Apps (Published via projects)	33
Public Shiny-based Apps (PUBLIC_WEB_APP)	15

Slide 36

- RS1** Include a URL.
Renee Sears, 2026-07-21T04:38:27.042
- RS2** Include a URL
Renee Sears, 2026-07-21T04:38:43.620

Get The Support You Need

Projects
Data
Public Apps
Public Projects
Developer
cruthfisher3

Dashboard
Files
Apps
Tasks
Data Studio
Troubleshooting Failed Tasks
Interactive Browsers

FAILED Task 1 - StringTie run
[Get support](#)
[View stats & logs](#)

Executed on June 12, 2020 07:53 by sevenbridges

Spot Instances: **On** | Memoization (WorkReuse): **Off** | Price: \$0.01 | Duration: 5 minutes

App: StringTie - Revision: 0

Error:
This task ran into a problem during execution and did not finish.
[Show details](#)

Inputs

App Settings
Show all

Output Settings

Aligned reads

HCC1143-CCLE-RNASeq-subset01.genome_aligned...

Reference annotation file

Homo_sapiens.GRCh38.84.gtf

Create input files for Ballgown and DESeq2
No value

Disable trimming
True

Ignore alignments on the specified sequence
No value

Keep annotated transcripts only
No value

Maximum fraction of multiply mapped reads
0.95

Minimum anchor length for junctions
10

Minimum isoform abundance
0.1

Minimum isoform length
200

Minimum junction coverage
1

Minimum locus gap separation value
50

Minimum read coverage
2.5

Number of threads
2

Output covered reference transcripts
False

Output gene abundance
False

Transcripts name prefix
STRG

Archived ballgown input tables
No value

Assembled transcripts
No value

Covered reference transcripts
No value

DESeq2 gene count matrix
No value

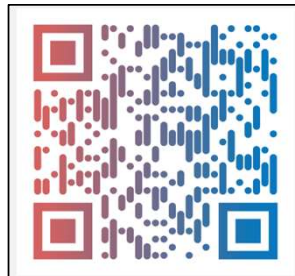
DESeq2 transcript count matrix
No value

Gene abundance estimation
No value

Slide 37

RS1 Highlight the get Support and the (?) or blow those sections out on the right.
Renee Sears, 2026-07-21T04:39:30.199

Get The Support You Need



Every Week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday

NCI Funding Is Available on the CGC

Pilot Credit Funds

- **\$300 of cloud credits** RS2
- Free for new CGC users
- Easy to request when signing up
- Fast approval



Collaborative Projects

- Cost estimation, optimization, and planning support
- Great for researchers new to bioinformatics and cloud approaches
- **Up to \$10k compute/storage costs**
- Fast, rolling applications
- To date > 60 projects

Slide 39

- RS1** Include a URL.
Renee Sears, 2026-07-21T04:38:27.042
- RS2** Include a URL
Renee Sears, 2026-07-21T04:38:43.620
- JC2 0** I'm not sure this is true anymore!!
Jeni Cliffe, 2026-07-21T14:52:50.587



Support@Velsera.Com

<https://meet.google.com/kbs-ojnj-dcg>



Every Week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday

VELSERA

Stay in touch.



Jeni Cliffe

Community Engagement Manager
Jenepher.Cliffe@velsera.com



Renee Sears, PhD

Community Engagement Manager
Renee.Sears@velsera.com



Michelle Engle, PhD

Community Engagement Manager
Michelle.Engle@velsera.com

RS1

Slide 40

RS1 Needs updates, but also IDK if my email should be on here given I don't really support CGC.
Renee Sears, 2026-07-21T04:40:30.824

Thank you.♦

Learn more at
CancerGenomicsCloud.org

VELSERA