

The power of connection

How the Cancer Research Data Commons enables researchers to connect data, computational tools, and collaborators to accelerate discovery

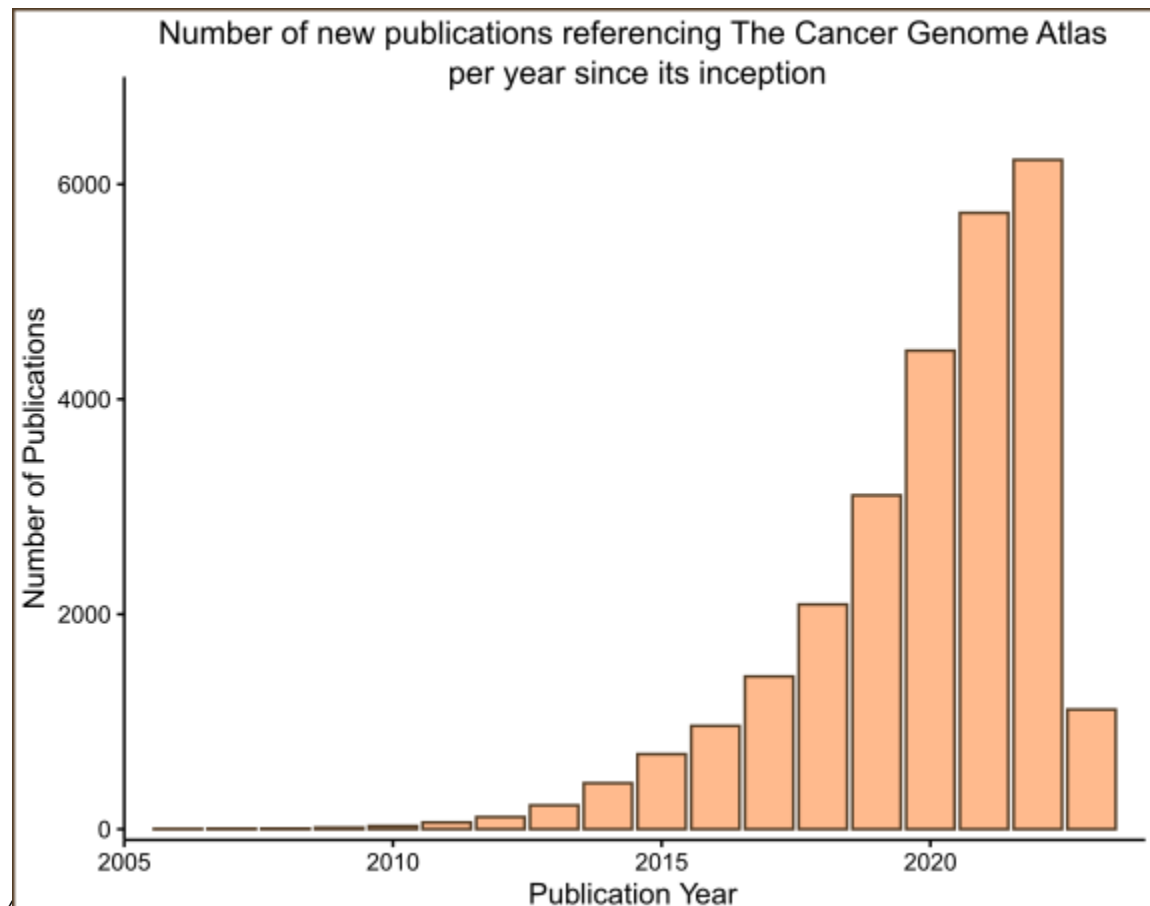
Brandi Davis-Dusenbery, PhD

Bioinformatics Training & Education Program (BTEP)

May 4, 2023



NCI has proven that large, shared data repositories accelerate discovery and improve patient outcomes.



March 4, 2013

[← BACK TO PRESS RELEASES](#)

Foundation Medicine Receives Accreditation from the College of American Pathologists

Jan 22, 2019

The FDA grants breakthrough device designation for TruSight Assay

Based on the content of TruSight Oncology 500, the proposed in vitro diagnostic will receive prioritized review and resources

Feb 16, 2023

Medicare Extends Coverage of Natera's Signatera™ MRD Test to Breast Cancer

Coverage to include serial monitoring in all subtypes, including hormone receptor-positive, HER2-positive, and triple negative breast cancers

NATIONAL CANCER INSTITUTE
Genomic Data Commons

Home Projects Exploration Analysis Repository

Quick Search Manage Sets

Human Tumor Atlas Network

Adrenal Gland
 Bile Duct
 Bladder
 Bone
 Bone Marrow
 Brain
 Breast
 Cervix
 Colorectal
 Esophagus
 Eye
 Head and Neck
 Kidney
 Liver
 Lung
 Lymph Nodes
 Nervous System
 Ovary
 Pancreas
 Pleura
 Prostate
 Skin
 Soft Tissue
 Stomach
 Testis
 Thyroid
 Uterus

Breast, TCGA-BLCA, TCGA-A5-A0G2

Summary Data Release 37.0 - March 29, 2023

PRIMARY SITES: 68
 CASES: 86,962
 GENES: 22,501
 MUTATIONS: 2,885,293

NATIONAL CANCER INSTITUTE
Genomic Data Commons

Explore Images Collections Getting Started User Forum News About

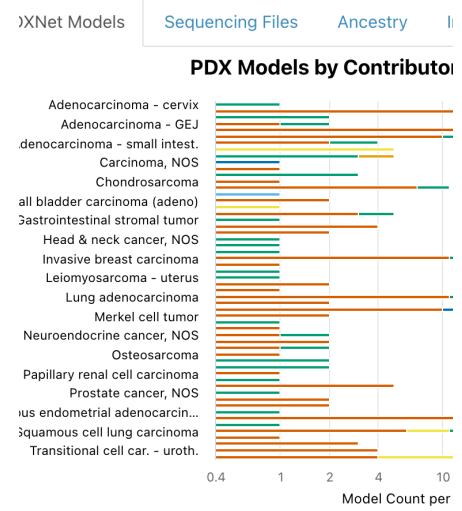
Positron Emission Tomography (PET)

Human Tumor Atlas Network

HTAN is a National Cancer Institute (NCI)-funded Cancer MoonshotSM initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease. (*Cell April 2020*)

Explore latest Data Learn more about HTAN

Today, TCGA is only the tip of the iceberg.



NIH NATIONAL CANCER INSTITUTE
Proteomic Data Commons

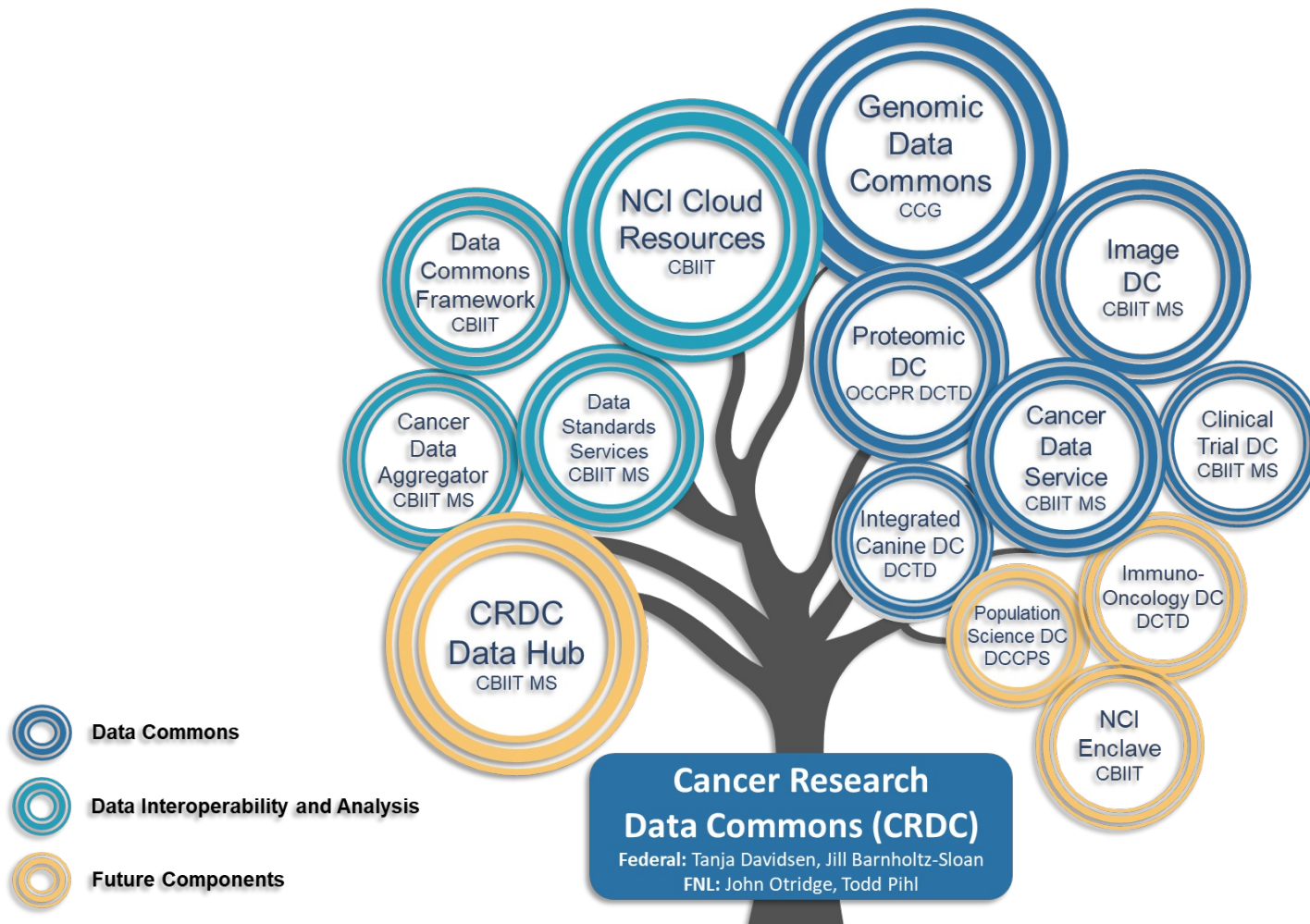
HOME

124 Studies 35 TB Data volume 111,713 Data files

Cases by Major Primary site

Primary Site	Number of Cases
Bone Marrow	~100
Brain	~100
Breast	~100
Colorectal	~100
Head and Neck	~100
Kidney	~100
Liver	~100
Lung	~100
Ovary	~100
Pancreas	~100
Stomach	~100
Uterus	~100

The NCI Cancer Research Data Commons connects diverse data with analytical power and tools.





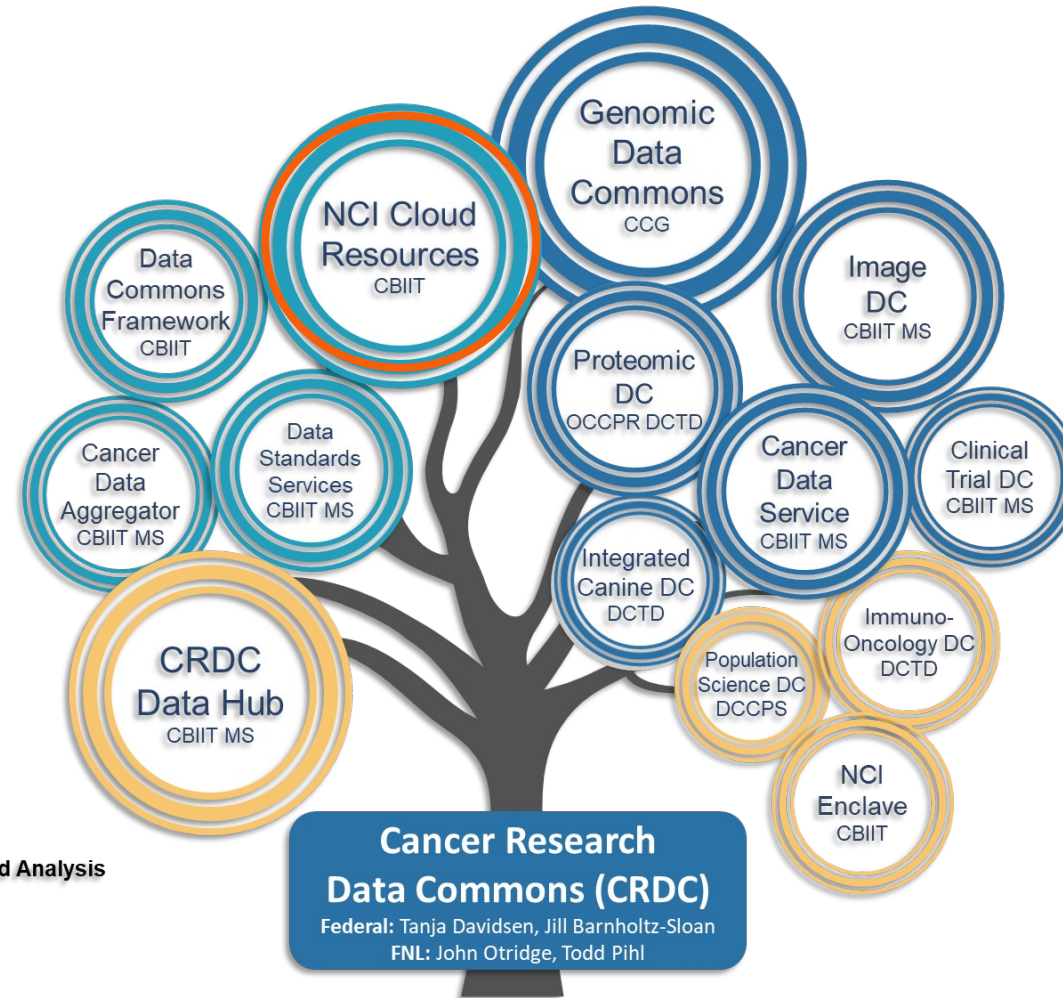
CANCER GENOMICS CLOUD

SEVEN BRIDGES

3+	1600+	800+	8000+	80000
Petabytes Public Data	Years of Compute	Public Tools & workflows	Users	User-created workflows

Provides 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



- Data Commons
- Data Interoperability and Analysis
- Future Components

*Data Hub, CDA, & CTDC have not reached go-live

Who are the CGC Users?

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



ADMINISTRATORS

- Manage and Control Users
- Monitor and Control Institutional Assets
- Manage and Monitor Projects
- Monitor and Control Costs
- Create Reports



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 and Gather Insights
- Easily Collaborate with Other 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



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 with Upstream/ Downstream Systems

Analyzing the genomic landscape of CHIP in solid tumor microenvironments.

Rutgers PhD Candidate Vaidhyathan Mahaganapathy



Data Connectivity

- Immediate access to TCGA data without needing to deal with long term storage.
- Originally proposed 5 tumor types, further optimization of tools allowed to extend to all qualifying tumor types.



Speed and Scale

- Ran parallelized analysis over a weekend that would have taken 3+ months on HPC.
- Created millions of files over the course of the analysis which also excluded use of local resources.



Community

"Office hours is a huge benefit, not only to help me, but also to not feel like I'm in a silo by myself."

- Optimized MERIT to be made available in tool repository to enable others to extend analysis.

Accelerating collaboration, data harmonization, and sharing for large consortia.

Patient Derived Xenograft Network



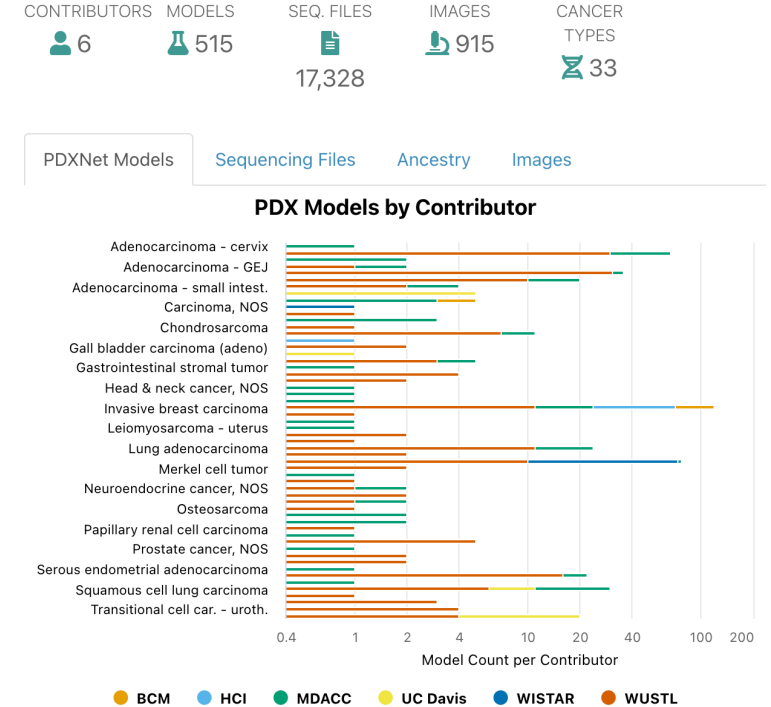
Data Connectivity

- Multimodal data including: WXS, RNAseq, proteomics and Imaging
- 16 optimized workflows published in CWL.
- Cost predictor based on 7,000 observations



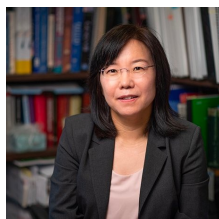
Impact

- 60+ manuscripts published by PDXNet investigators over 5 years.
- Preclinical, *in vivo* evidence supporting 10 Clinical Trial LOIs incl ComboMatch & ETCTN



Promoting bioinformatics education for the next generation of cancer researchers.

Diverse approaches and engagement strategies tailored to community needs



Train the trainer

Min Zhang, M.D., Ph.D., researcher within the Department of Statistics at Purdue University

Developed and led an NCI-funded R25 initiative called "Big Data Training for Cancer Researchers".

After the 2022 workshop, **77% of the participants stated in a survey that they were interested in or already using the Cancer Genomics Cloud** to conduct their own research.

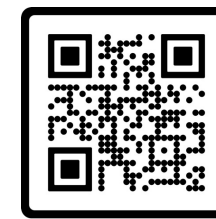


Workshops

Judy Wawira Gichoya, MS, MD | National Institute of Biomedical Imaging and Bioengineering

Over **200 international scientists** and stakeholders from across Africa as part of DS-I Africa initiative.

Workshop on data harmonization using cloud resources, then a live demonstration showcasing a deep learning AI tool utilizing a Jupyter Labs Python Notebook to make predictions on one of the Covid-19 imaging datasets provided by Judy



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

Introduction to the CGC

Build and Execute Flexible Analytical Pipelines



User friendly portal

Access petabytes of publicly available data, and analyze it alongside private data.



Unprecedented Collaboration Features

Collaborate within and across organizations while keeping control of your assets with precise permission levels.



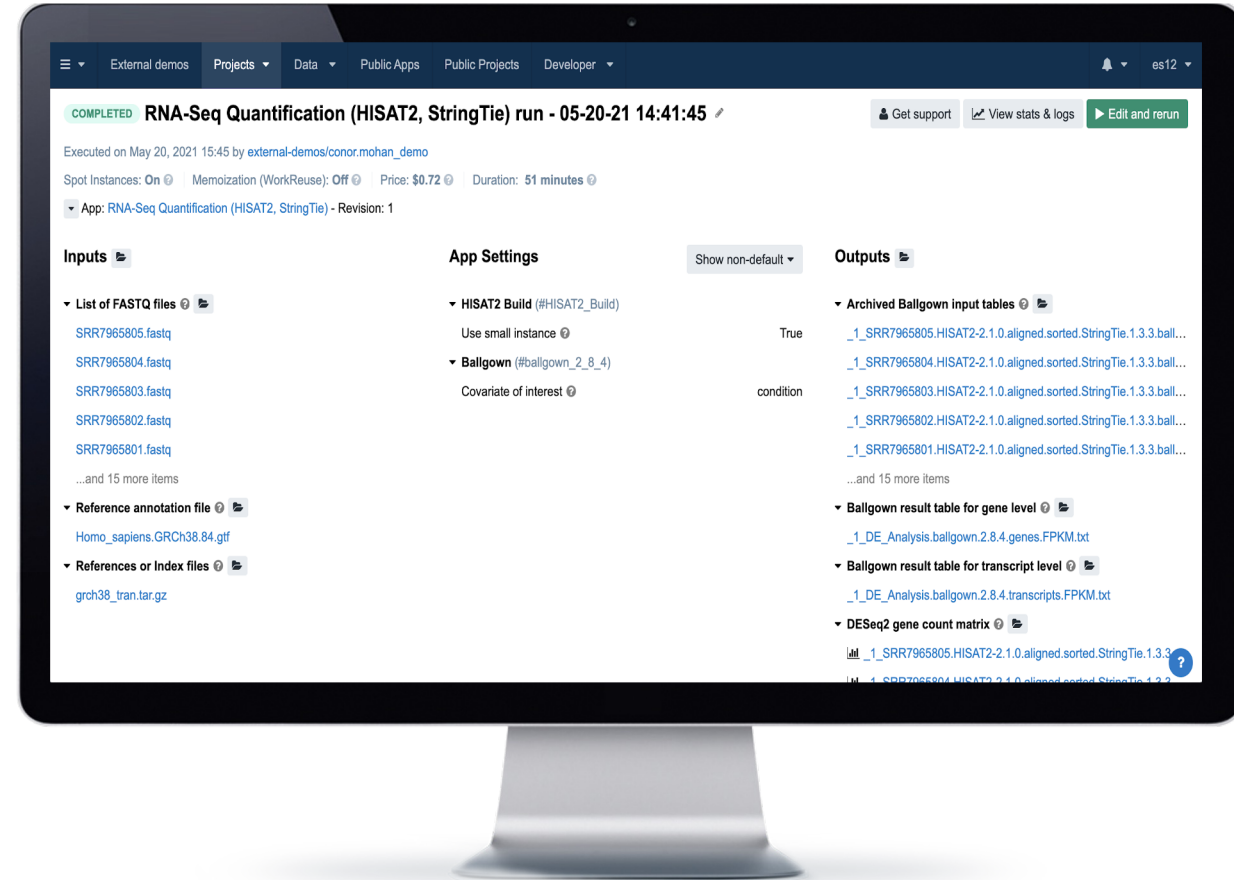
Industry Standard Tools: Reproducibility

Execute, build, and customize analysis pipelines using popular tools such as CWL, WDL, Nextflow, Docker, RESTful APIs, and CLI.



Connected Cloud

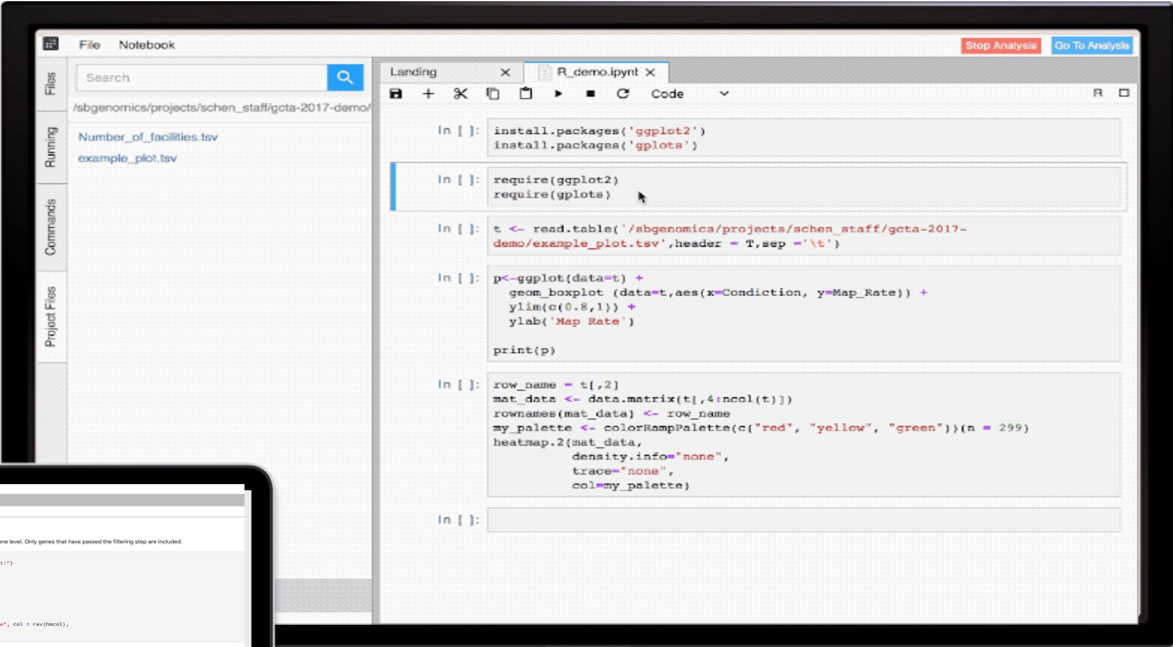
Keep data in your own buckets with storage support for AWS, and Google Cloud Platform.



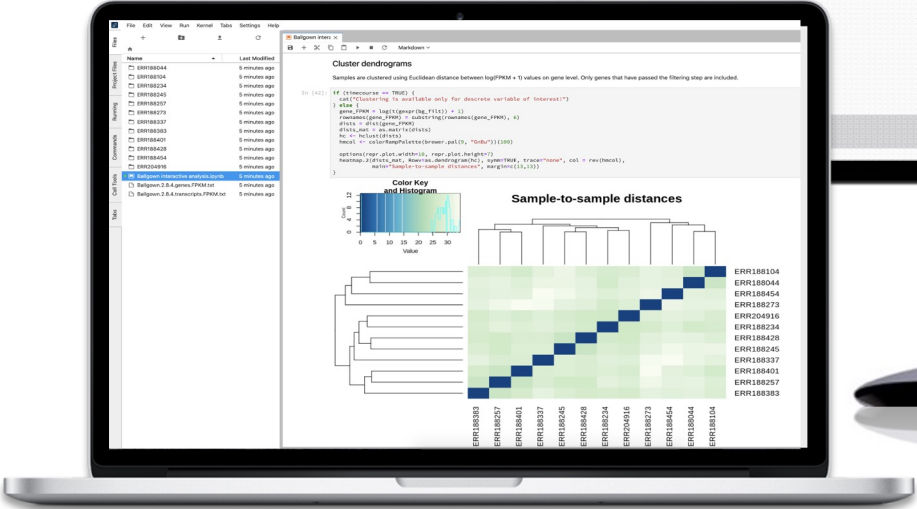
Integrated Custom Tertiary Analysis Tools

Data Science Workbench

Derive new insights using interactive analysis environments with JupyterLab, and RStudio environments. Code in Python and R, and create Jupyter Notebooks to record and share your analyses.



COMING SOON



Galaxy PROJECT

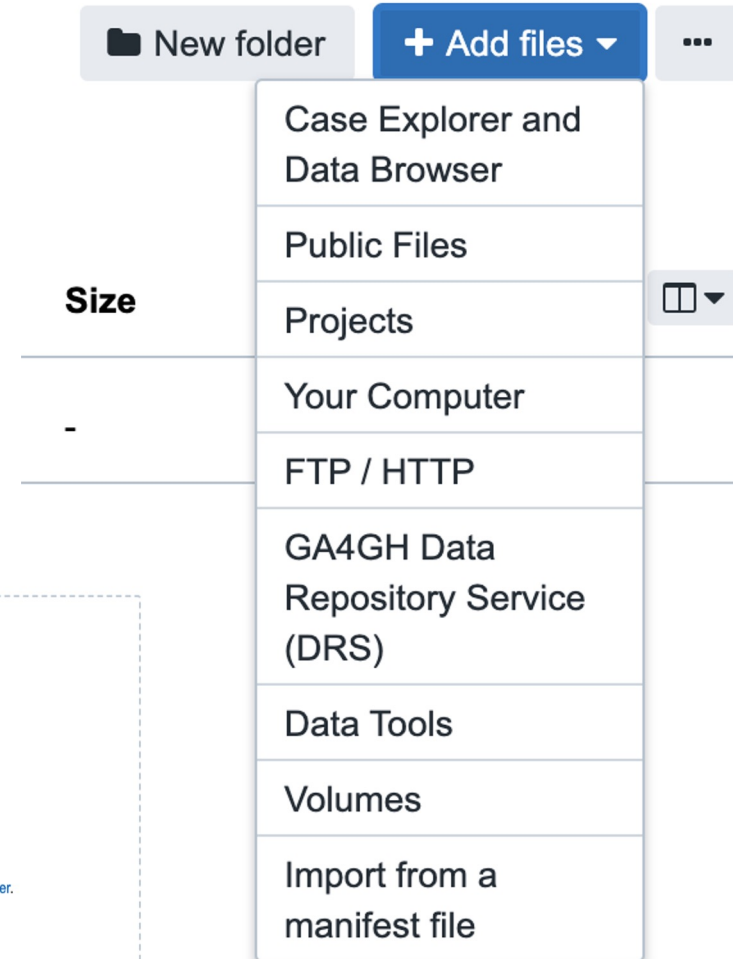
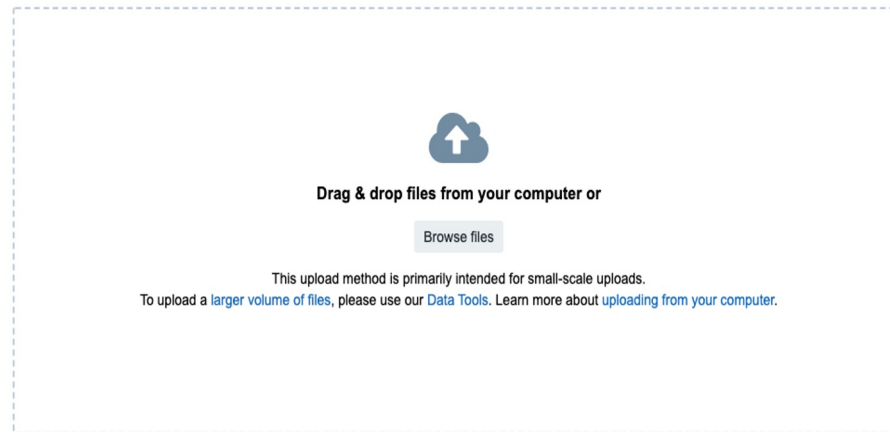
sas

VELSERA

Petabytes of public data, at your fingertips



**Upload your data,
using the user
interface portal or the
command line**



Making an impact

112

Publications + Citations

30+

Webinars + Video tutorials

100+

Office hour sessions

VELSERA

LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint

Wilson McKerrrow^{a,b}, Xuya Wang^{a,b}, Carlos Mendez-Dorantes^{c,d}, Paolo Mita^{a,b}, Song Cao^{a,f}, Mark Grivainis^{a,b}, Li Ding^{a,f}, John LaCava^{a,h}, Kathleen H. Burns^{c,d}, Jef D. Boeke^{a,b,i,1}, and David Fenyö^{a,b,1}

^aInstitute for Systems Genetics, New York University Grossman School of Medicine, New York, NY 10016; ^bDepartment of Biochemistry Pharmacology, New York University Grossman School of Medicine, New York, NY 10016; ^cDepartment of Oncologic Pathology, Boston, MA 02215; ^dDepartment of Pathology, Harvard Medical School, Boston, MA 02115; ^eDepartment of Medicine and Gene Washington University in St. Louis, St. Louis, MO 63110; ^fMcDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63110; ^gCellular and Structural Biology, The Rockefeller University, New York, NY 10065; ^hEuropean Research Institute for the Biology of Center Groningen, 9713 GZ Groningen, The Netherlands; and ⁱDepartment of Biomedical Engineering, Tandon School of Engineering, New York University, New York, NY 10002

Contributed by Jef D. Boeke; received August 31, 2021; accepted January 14, 2022; reviewed by Molly Gale Hammell and Michael

RNA BIOLOGY
2022, VOL. 19, NO. 1, 279–289
<https://doi.org/10.1080/15476286.2022.2025680>

RESEARCH PAPER

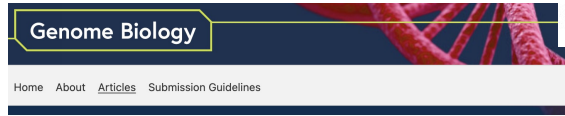
Flexible pri-miRNA structures enable tunable production of 5' isomiRs

Xavier Boffill-De Ros^a, Zhenyi Hong^a, Ben Birkenfeld^a, Sarangelica Alamo-Ortiz^a, Acong Yang^a, Lisheng Dai^a, and Shuo Gu^a

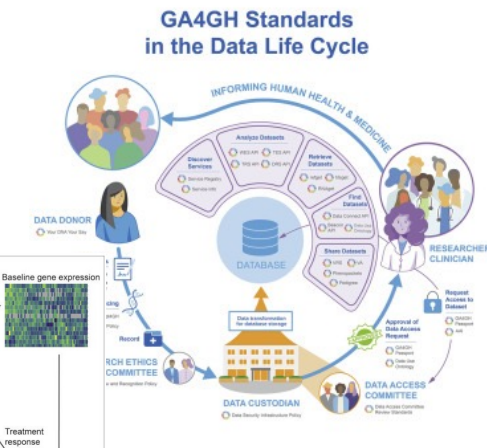
^aRNA Mediated Gene Regulation Section, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD, USA; ^bNeural Development Section, Mouse Cancer Genetics Program, Center for Cancer Research, National Cancer Institute, Frederick, MD, USA



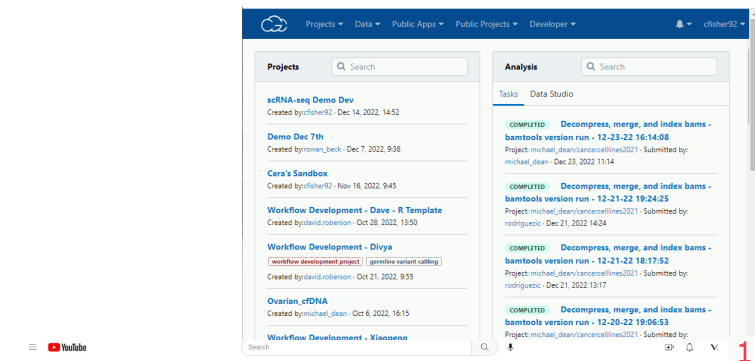
OPEN ACCESS Check for updates



Short Report | Open Access | Published: 05 August 2021
Specific splice junction detection in single cells with SICILIAN
Roozbeh Dehghanasari, Julia Eve Olivieri, Ana Damjanovic & Julia Salzman
Genome Biology 22, Article number: 219 (2021) | Cite this article
1336 Accesses | 30 Altmetric | Metrics



Register here:
bit.ly/CGCwebinar



Live demo!!



Trends in Cancer Supports open access Submit

TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines

Ashkan Shahbandi • Hoang D. Nguyen • James G. Jackson  

Published: February 05, 2020 • DOI: <https://doi.org/10.1016/j.trecan.2020.01.007> •  Check for updates

What gene expression changes are present in breast cancer tumors harboring p53 nonsense mutation compared to breast cancer tumors which express p53 at a similar expression level without mutation.

lights

words

ferences

le info

ted Articles

Highlights

TP53 is mutant in 30% of all breast cancers and is the most frequently mutated gene. The role of *TP53* in the management of breast cancer remains unclear.

Recent analysis reveals that mutant p53 can be detrimental or beneficial to clinical outcome, depending on treatments given.

Live demo!!

Differential expression analysis

Table of adjusted p-values ordered features:

##	target_id	ext_gene	num_aggregated_transcripts	sum_mean_obs_counts	pval	qval
## 1	ENSG00000104419	NDRG1	16	62.18210	7.953447e-10	2.238339e-05
## 2	ENSG00000073849	ST6GAL1	12	42.41697	5.085899e-09	7.156623e-05
## 3	ENSG00000126217	MCF2L	19	78.84039	1.554294e-08	1.458083e-04
## 4	ENSG00000141552	ANAPC11	19	72.25131	3.160626e-08	2.223737e-04
## 5	ENSG00000242265	PEG10	5	28.77247	8.146654e-08	4.585426e-04
## 6	ENSG00000099204	ABLIM1	6	26.49753	1.930873e-07	9.056759e-04

> J Natl Cancer Inst. 2022 Apr 11;114(4):579-591. doi: 10.1093/jnci/djab222.

NDRG1 in Aggressive Breast Cancer Progression and Brain Metastasis

Emily S Villodre ^{1 2}, Xiaoding Hu ^{1 2}, Bedrich L Eckhardt ^{1 2 3}, Richard Larson ^{2 4}, Lei Huo ^{2 5}, Ester C Yoon ⁵, Yun Gong ^{2 5}, Juhee Song ⁶, Shuying Liu ¹, Naoto T Ueno ^{1 2}, Savitri Krishnamurthy ^{2 5}, Stefan Pusch ^{7 8}, Debu Tripathy ^{1 2}, Wendy A Woodward ^{2 4}, Bisrat G Debeb ^{1 2}

Affiliations + expand

PMID: 34893874 PMID: PMC9002276 DOI: 10.1093/jnci/djab222 

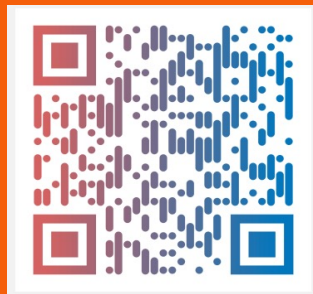
[Free PMC article](#)

Abstract

Background: N-Myc downstream regulated gene 1 (NDRG1) suppresses metastasis in many human malignancies, including breast cancer, yet has been associated with worse survival in patients with inflammatory breast cancer. The role of NDRG1 in the pathobiology of aggressive breast cancers remains elusive.



Create an account and
access free credits at
CancerGenomicsCloud.org



Every Week:

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

Stay in touch.



Cera Fisher, PhD

Community Engagement Manger
Cera.Fisher@velsera.com



Rowan Beck, PhD

Community Engagement Manager
Rowan.Beck@velsera.com



Zelia Worman, PhD

Director of Researcher Engagement and
Education
Zelia.Worman@velsera.com