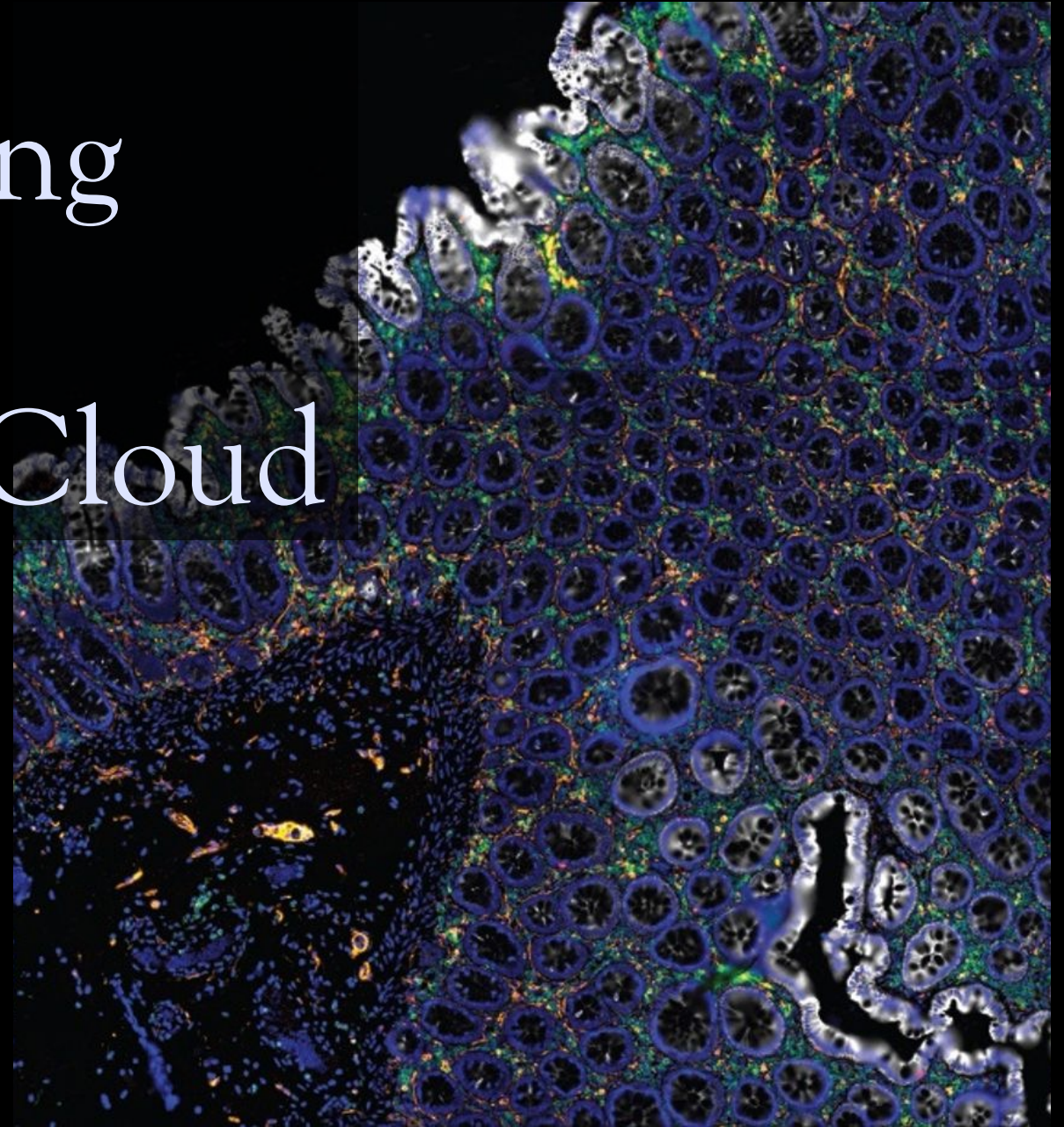


Bioinformatics Using The Seven Bridges Cancer Genomics Cloud

Seven Bridges Platform

Rowan Beck, PhD
March 27th, 2024

VELSERA



Agenda

Overview of the Seven Bridges Cancer Genomics Cloud, powered by Velsera
Features of the CGC
Live Demonstration



CANCER GENOMICS CLOUD

SEVEN BRIDGES

3+

Petabytes
Public Data

1600+

Years of
Compute

900+

Public Tools
& Workflows

8000+

Users

80000+

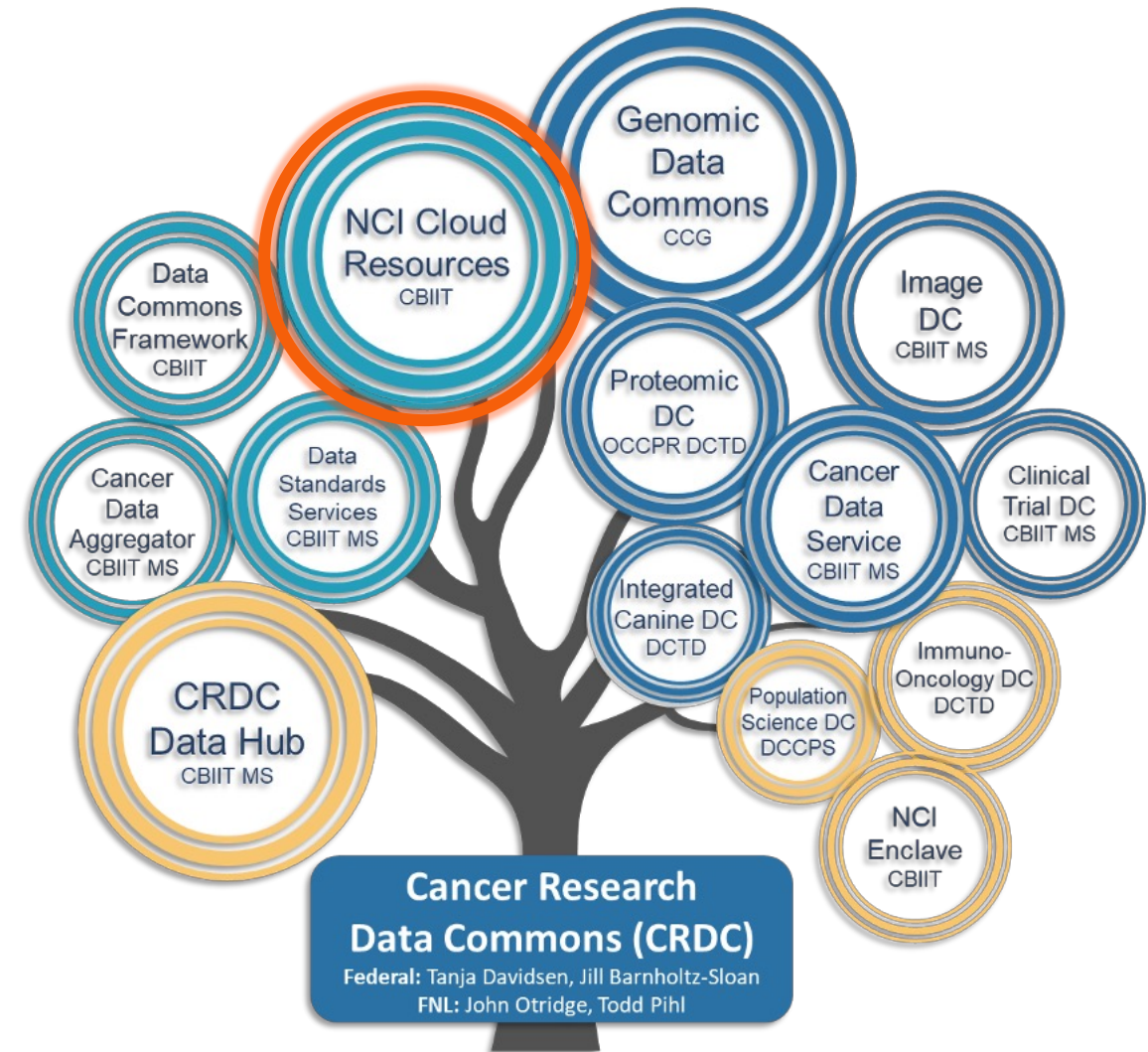
User-Created
Workflows

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

VELSERA



Data Commons



Data Interoperability and Analysis



Future Components

Who are the 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 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



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 with Upstream/ Downstream Systems

Features of the CGC

Access Public Data, or Use Your Own

The screenshot shows the Velsera Data Studio interface. The top navigation bar includes 'Projects', 'Data', 'Public Apps', 'Public Projects', and 'Developer'. The main header displays 'Demonstration: Building an App' and 'Interactive Browsers'. The left sidebar shows 'Files' with a search bar and filter options for 'Extension', 'Sample ID', 'Task ID', and 'Tags'. The main content area contains a table of files:

Name	Task ID	Created on	Extension	Size
G20479.HCC1143.2_1Mreads.tar.gz <small>(TEST)</small>	-	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

The 'Add files' dropdown menu is open, showing options: 'Case Explorer and Data Browser', 'Public Files', 'Projects', 'Your Computer', 'FTP / HTTP', 'GA4GH Data Repository Service (DRS)', 'Data Tools', 'Volumes', and 'Import from a manifest file'.



Browse Hundreds of Tools and Workflows

Public apps for your data analysis

We offer publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

[Browse 939 apps](#)

GRAF Germline Variant Detection Workflow

The GRAF Germline Variant Detection Workflow enables accurate alignment and variant calling by utilizing a genome graph reference that can address the bias and other limitations inherent in linear genome references. Seven Bridges has constructed a comprehensive pan-genome graph that incorporates the...

[Alignment](#) [Variant Calling](#) [Graph](#)

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

Public Apps

MCMICRO

Search result: 1 Item

MCMICRO

MCMICRO is an end-to-end processing pipeline for multiplexed whole slide imaging and tissue microarrays. It comprises ...

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

MCMICRO

Enables the processing of multiplexed tissue images.

Transform whole-slide images into single-cell data using this simple workflow.

No Coding Required to Run an Analysis

The screenshot displays the VELSERA web interface for configuring a task. The top navigation bar includes 'Projects', 'Data', 'Public Apps', 'Public Projects', and 'Developer'. The user 'rowan_beck_era' is logged in. The task title is 'DRAFT Differential Expression - Salmon + DESeq2 run - 11-30-23 17:12:19'. The task is currently in a 'DRAFT' state, and there are buttons for 'Get support', 'Discard', and 'Run'. The task was last updated by 'rowan_beck_era' on Nov. 30, 2023 12:12. The application being used is 'Differential Expression - Salmon + DESeq2 - Revision: 1'. The 'Task Inputs' tab is active, showing three main sections: 'Inputs', 'App Settings', and 'Output Settings'. The 'Inputs' section includes 'Batching' (Off), 'FASTQ read files' (5 files selected), 'GTF annotation' (1 file selected), and 'Genome FASTA' (No files selected). The 'App Settings' section includes 'DESeq2' parameters: 'Covariate of interest' (Genotype), 'Factor level - reference' (WT), and 'Factor level - test' (KD). The 'Output Settings' section lists various output options, all of which are currently set to 'No value'. An 'Activity Monitor' button is located at the bottom of the configuration area.

Inputs

Batching Off

FASTQ read files *

- SRR9058997_1.fastq
- SRR9058993_1.fastq
- SRR9058992_2.fastq
- SRR9058992_1.fastq
- SRR9058991_2.fastq
- ...and 25 more items

GTF annotation *

- GRCh38ERCC.ensembl95.gtf

Genome FASTA

No files selected

App Settings

DESeq2 (#deseq2_1_26_0)

Covariate of interest *

Genotype

Factor level - reference

WT

Factor level - test

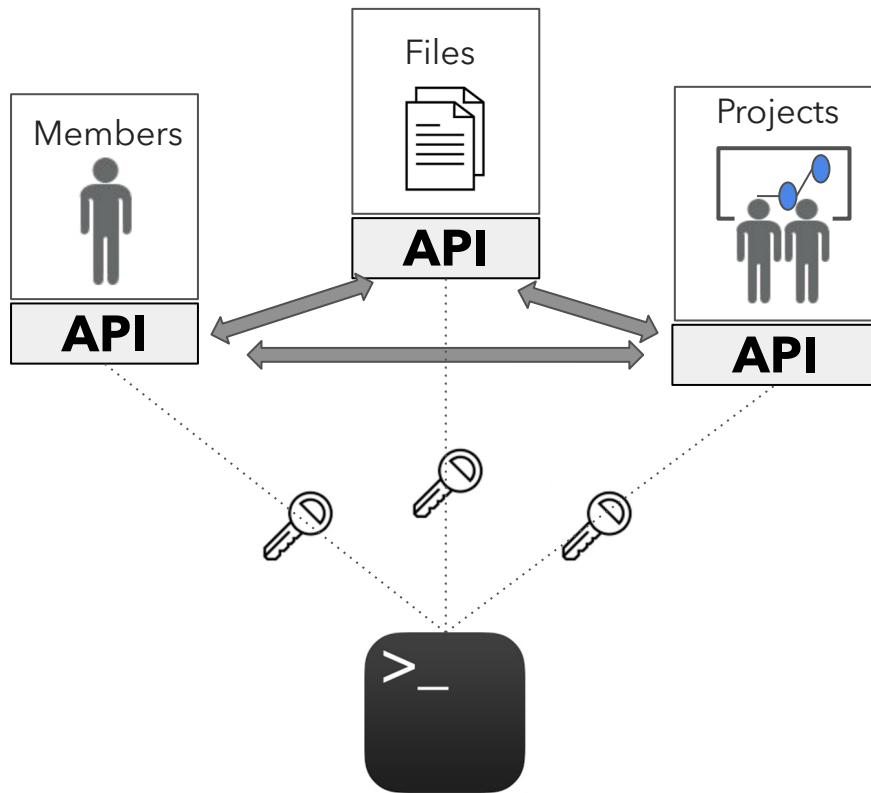
KD

Output Settings

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

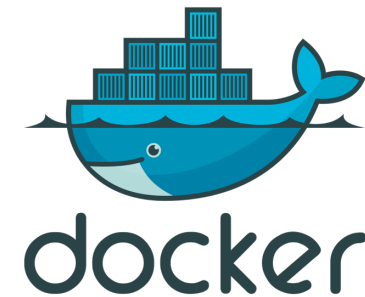
Activity Monitor

Coding Features



Example platform with Service Oriented Architecture and API access

- REST API
- API Bindings in Python, R, and Java

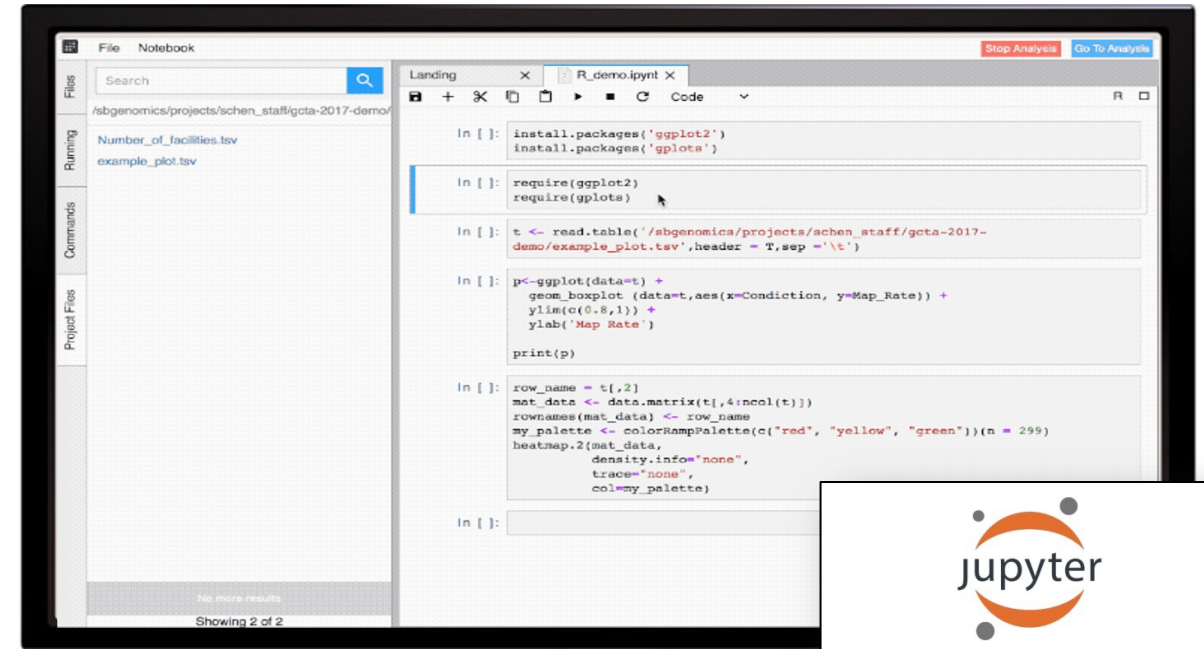


Integrated Custom Tertiary Analysis Tools

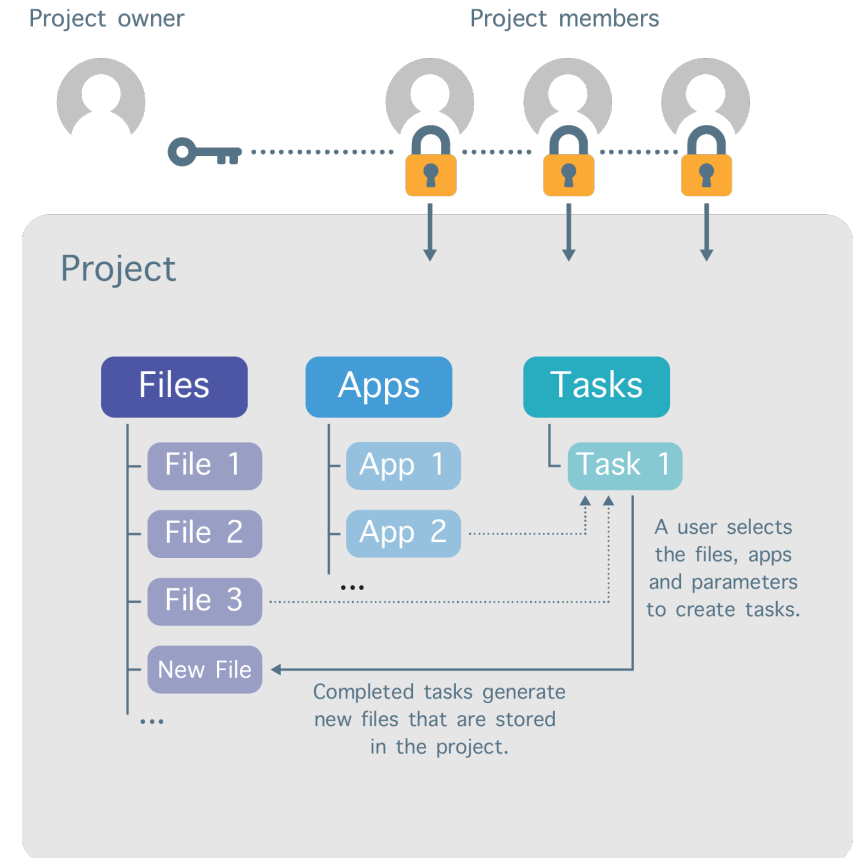
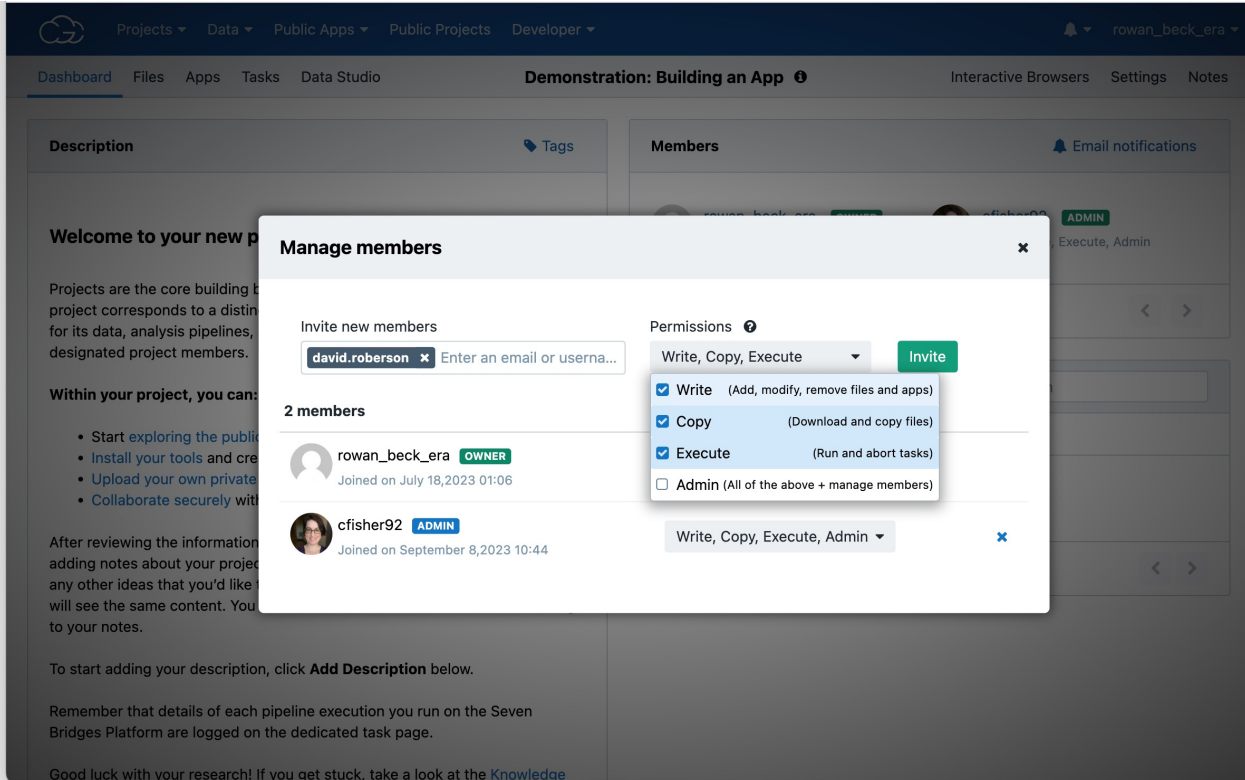
Data Science Workbench

Derive new insights using interactive analysis environments with JupyterLab, RStudio, Galaxy, and SAS Studio environments. Create scripts and notebooks to record and share your analyses.

OHIF Image Viewer allows users to easily interact with DICOM images.



Collaborating Has Never Been Easier



Collaborating Has Never Been Easier

Pre-loaded with

- input and output files
- pre-run tasks
- recommended settings

Detailed descriptions on what the workflow does and how to use the pipeline

The screenshot displays the Velsera interface for the 'MCMICRO - End to End Microscopy Image Processing' project. The top navigation bar includes 'Projects', 'Data', 'Public Apps', 'Public Projects' (highlighted with an orange box and a red arrow), and 'Developer'. Below the navigation, the page is divided into two main sections: 'Description' and 'Analysis'.

Description: The title is 'MCMICRO - End to End Microscopy Image Processing Public Project'. The text describes MCMICRO as an end-to-end processing pipeline for multiplexed whole slide imaging and tissue microarrays. It mentions that this is a CWL wrapper designed by Seven Bridges. The description also notes that the project demonstrates the usage of two CWL versions of MCMICRO, namely 'MCMICRO' and 'MCMICRO for HTAN'.

Analysis: This section shows a list of completed analysis runs. Each entry includes a 'COMPLETED' status, the run name, and the submission date and time. The runs listed are:

- MCMICRO run - Exemplar 002 (Submitted by: sevenbridges · Apr 13, 2023 16:16)
- MCMICRO run - Exemplar 001 (Submitted by: sevenbridges · Apr 13, 2023 16:13)
- MCMICRO for HTAN run - WD-76845 (Submitted by: sevenbridges · Apr 13, 2023 15:59)
- MCMICRO for HTAN run - HTMA402 (Submitted by: sevenbridges · Apr 13, 2023 15:56)

Citation: The section provides information on how to cite the project. It states that project participants agree to acknowledge the funding for the CGC in all publications and external presentations. The citation text is: "The Seven Bridges Cancer Research Data Commons Cloud Resource has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I and 75N91019D00024." It also provides the citation for Lau et al (2017) The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. *Cancer Res.* 77(21):e3-e6. doi: 10.1158/0008-5472.CAN-17-0387.

Estimate Cloud Costs

Performance Benchmarking

- Runtimes
- Task Costs
- Various file sizes

Experiment type	Input size	Paired-end	# of reads	Read length	Duration	Cost	Instance (AWS)
RNA-Seq	2 x 230 MB	Yes	1M	101	18min	\$0.40	c4.8xlarge
RNA-Seq	2 x 4.5 GB	Yes	20M	101	30min	\$0.60	c4.8xlarge
RNA-Seq	2 x 17.4 GB	Yes	76M	101	64min	\$1.20	c4.8xlarge

Cloud Cost Estimator

- Available for a limited number of apps
- Create an estimate for the cost of your specific use case
- Compare workflows to save on computing costs

The screenshot shows the Cloud Cost Estimator interface. At the top, there is a navigation bar with 'Projects', 'Data', 'Public Apps', and user information. A search bar and filters for 'Type', 'Category', and 'Toolkit' are visible. A dropdown menu shows 'CWL version: v1.0 +2' and 'Cost Estimator: Available'. A modal window titled 'Cost estimator: STAR' is open, displaying the following information:

Cost estimator: STAR

The cost is estimated based on these parameters:

- Spot Instances**
On
- File size**
119.75 GB
- Instance type**
Default

Cost Estimation **\$0.32 - 0.58**

NCI Funding Is Available on the CGC

Diverse approaches and engagement strategies tailored to community needs

Pilot Credit Funds

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

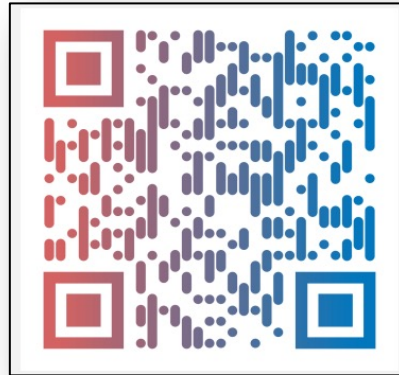


Collaborative Projects

- 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



Get The Support You Need



Every Week:

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

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

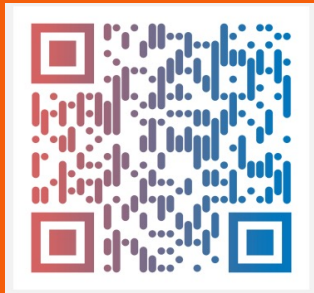
App: StringTie - Revision: 0

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

Inputs	App Settings	Output Settings
Aligned reads HCC1143-CCLE-RNASeq-subset01.genome_aligned...	Create input files for Ballgown and DESeq2: No value	Archived ballgown input tables: No value
Reference annotation file Homo_sapiens.GRCh38.84.gtf	Disable trimming: True	Assembled transcripts: No value
	Ignore alignments on the specified sequence: No value	Covered reference transcripts: No value
	Keep annotated transcripts only: No value	DESeq2 gene count matrix: No value
	Maximum fraction of multiply mapped reads: 0.95	DESeq2 transcript count matrix: No value
	Minimum anchor length for junctions: 10	Gene abundance estimation: No value
	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	



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 Manager

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

Thank you.

Learn more at
CancerGenomicsCloud.org

VELSERA