

# Introduction to the Cancer Genomics Cloud, powered by Seven Bridges

Zélia Worman, Ph.D.  
Program Manager





## Class Overview

- **Lecture 1 - Overview**  
Hands-on demo: your first project on the CGC
- **Lecture 2 - Bulk RNA-seq workflow(s) and sequence alignment**  
Hands-on demo: your CGC workflow
- **Lecture 3 - Differential expression and visualizations**  
Hands-on demo: your CGC visualization
- **Lecture 4 - Single-cell RNA-seq overview and workflow**  
Hands-on demo: Seurat



# Lecture Overview

- Overview of the Cancer Genomics Cloud, powered by Seven Bridges
- Highlight of key features of how to import data to the CGC and running an analysis
- Hands-on Demo – logging in setting up your first project

# Explosion of 'omics data with ease of sequencing

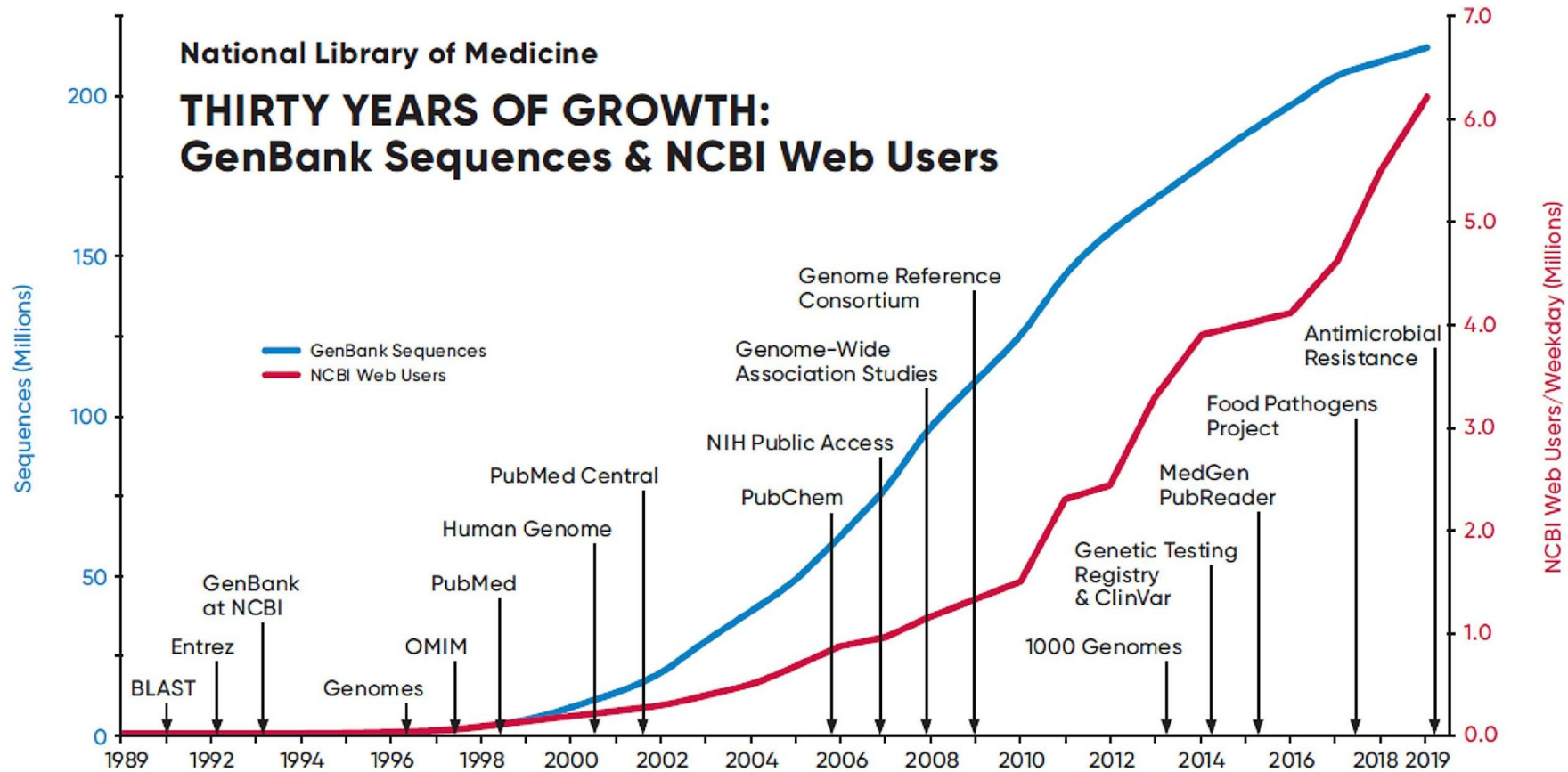


Fig. 1. **Growth of GenBank sequences and NCBI web users through 2019.** Figure from the Department of Health and Human Services National Institutes of Health. Jim Gaffney, et. al. Open access to genetic sequence data maximizes value to scientists, farmers, and society, Global Food Security, Volume 26, 2020.

# To give you one example, the TCGA is just one of many datasets currently available!

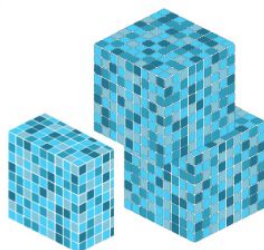
## NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

### TCGA BY THE NUMBERS

TCGA produced over

# 2.5

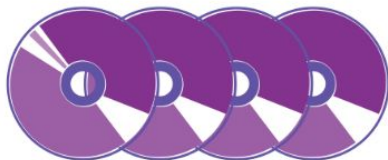
PETABYTES  
of data



To put this into perspective, **1 petabyte** of data is equal to

# 212,000

DVDs



TCGA data describes



# 33

DIFFERENT  
TUMOR TYPES

...including

# 10

RARE  
CANCERS

...based on paired tumor and normal tissue sets collected from



# 11,000

PATIENTS

...using

# 7

DIFFERENT  
DATA TYPES



## HOW BIG IS A PETABYTE?

### 11,000 4k movies



It would take you over 2.5 years of nonstop binge watching to get through a petabyte's worth of 4k movies



### 20+ PB of data in the Library of Congress



If you took a petabyte's worth of 1GB flash drives and lined them up end to end, they would stretch over

### 92 football fields



### 4,000 digital photos every day for the rest of your life

Sources: Lifewire.com,  
Blogs.loc.gov, cobaltiron.com

**cobalt IRON**



# Increasingly large datasets bring challenges to data analysis



# What is the cloud?



<https://blog.vsoftconsulting.com/blog/what-exactly-is-the-cloud>

# The CGC Democratizes Complex Analyses

- A stable, secure, and highly customizable cloud storage and computing platform
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines



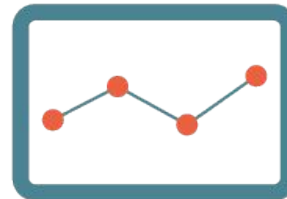
Easy data  
management



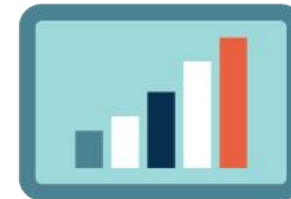
Secure  
collaboration &  
managed billing



Flexible & fully  
reproducible  
methods



Optimized  
bioinformatics  
algorithms



Scalable  
computation



Extensible &  
developer  
friendly tools

# The Seven Bridges Cancer Genomics Cloud (CGC)



LEARN



COLLABORATE



TEACH



INNOVATE

For more information, visit us at <https://www.cancergenomicscloud.org/>

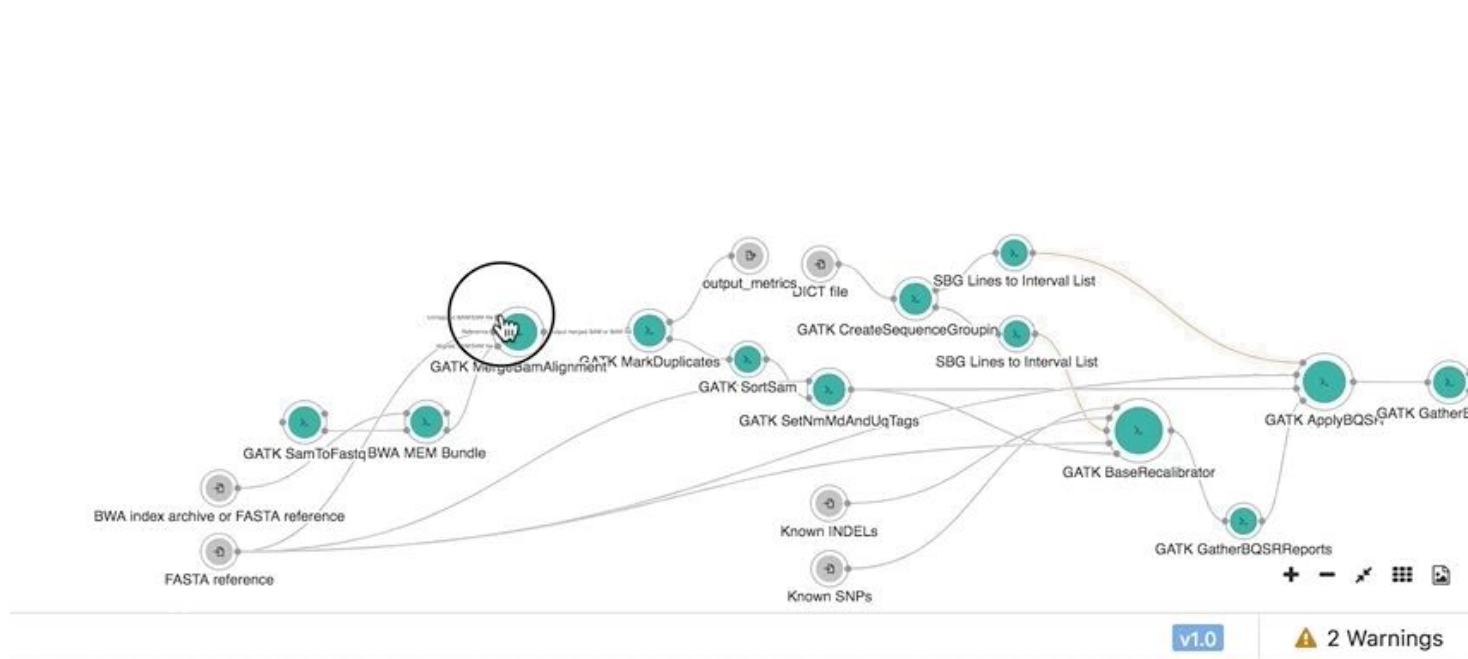
# The CGC provides an easy way to find data

Visually explore and access **3+ PB** of multi-omic public data through our user-friendly portal and APIs.

The screenshot displays the Cancer Genomics Cloud (CGC) user interface. At the top, a dark blue navigation bar contains menu items: 'Projects', 'Data', 'Public Apps', 'Public projects', and 'Developer'. On the right side of this bar, there is a user profile icon labeled 'zwamen'. Below the navigation bar, a secondary bar shows 'TCGA' and 'New query Edited'. To the right of this bar are buttons for 'Create new query', 'Queries', 'Search by ID', and 'Copy files to project'. The main content area features a red card with 'Case' and 'Investigation TCGA-BRCA' text. Below this, a summary bar shows 'Case 1,098' and buttons for 'Export', 'Details', and 'Analytics'. The 'Case' section is followed by an 'Investigation' section with a bar chart for 'TCGA-BRCA' showing a value of 1,098. At the bottom left, there are links for 'Forum', 'Terms', 'Privacy', and 'Data Use'. At the bottom right, there is a copyright notice '© 2021 Seven Bridges Genomics' and a help icon.

# The CGC provides an easy way analyze data

Use the **600+** cloud- and cost-optimized tools in our Public Apps library OR deploy custom tools using **Rabix Composer**, Jupyter notebooks or R packages



```
File Notebook Stop Analysis Go To Analysis  
Landing X R_demo.ipynb X  
/sbgenomics/projects/schen_staff/gcta-2017-demo/  
Running  
Number_of_facilities.tsv  
example_plot.tsv  
Commands  
Project Files  
In []: install.packages('ggplot2')  
install.packages('gplots')  
In []: require(ggplot2)  
require(gplots)  
In []: t <- read.table('/sbgenomics/projects/schen_staff/gcta-2017-  
demo/example_plot.tsv',header = T,sep = '\t')  
In []: p<-ggplot(data=t) +  
geom_boxplot(data=t,aes(x=Condition, y=Map_Rate)) +  
ylim(c(0.8,1)) +  
ylab('Map_Rate')  
print(p)  
In []: row_name = t[,2]  
mat_data <- data.matrix(t[,4:ncol(t)])  
rownames(mat_data) <- row_name  
my_palette <- colorRampPalette(c("red", "yellow", "green"))(n = 255)  
heatmap.2(mat_data,  
density.info="none",  
trace="none",  
col=my_palette)  
In []:
```



# High impact publications on the CGC

**PNAS** Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author, or

Home Articles Front Matter News Podcasts Authors

## RESEARCH ARTICLE



### Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers

#### Genome Biology

Home About Articles Submission Guidelines

Short Report | [Open Access](#) | [Published: 05 August 2021](#)

### Specific splice junction detection in single cells with SICILIAN

[Roozbeh Dehghannasiri](#), [Julia Eve Olivieri](#), [Ana Damjanovic](#) & [Julia Salzman](#)

[Genome Biology](#) **22**, Article number: 219 (2021) | [Cite this article](#)

1336 Accesses | 30 Altmetric | [Metrics](#)

**DATABASE**  
The Journal of Biological Databases and Curation

### Bioinformatics tools developed to support BioCompute Objects

[Janisha A Patel](#), [Dennis A Dean](#), [Charles Hadley King](#), [Nan Xiao](#), [Soner Koc](#), [Ekaterina Minina](#), [Anton Golikov](#), [Phillip Brooks](#), [Robel Kahsay](#), [Rahi Navelkar](#) ... [Show more](#)

*Database*, Volume 2021, 2021, baab008, <https://doi.org/10.1093/database/baab008>

**Published:** 30 March 2021 [Article history](#) ▼

**RNAbiology**  
Volume 18 | Issue 10 | 2021

### Novel, abundant Drosha isoforms are deficient in miRNA processing in cancer cells

[Lisheng Dai](#), [Lillian Hallmark](#), [Xavier Bofill De Ros](#) , [Howard Crouch](#), [Sean Chen](#), [Tony Shi](#), ...[show all](#)

Pages 1603-1612 | Received 02 Jan 2020, Accepted 19 Aug 2020, Accepted author version posted online: 20 Aug 2020, Published online: 30 Aug 2020

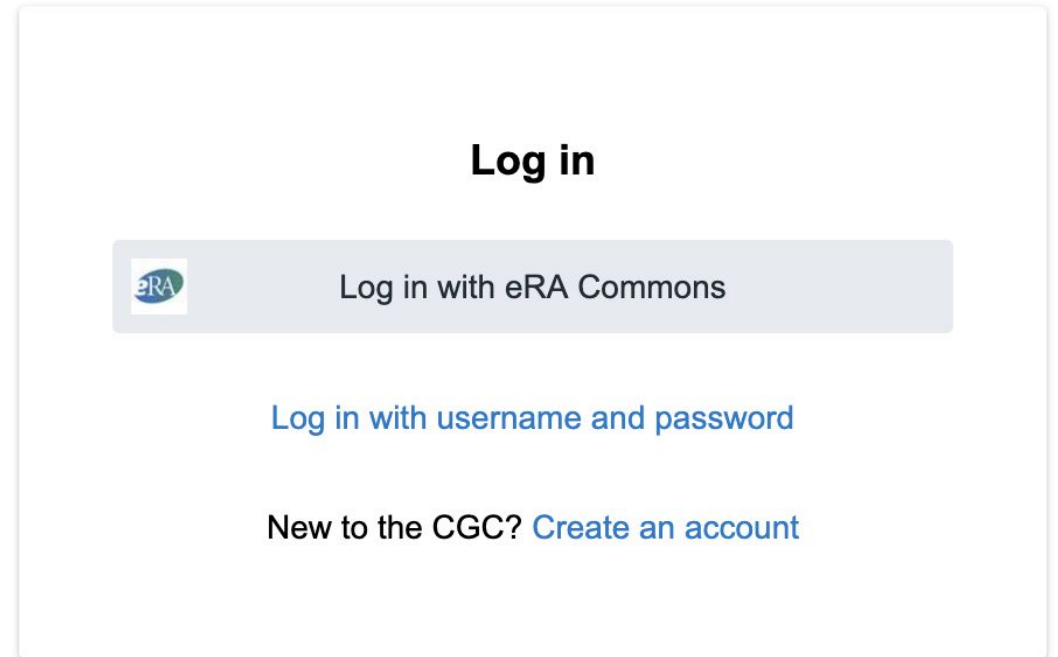
[Download citation](#) <https://doi.org/10.1080/15476286.2020.1813439>



# Quickly get an account on the CGC



- Sign up with your email
- Option to connect with eRA Commons to access controlled data
- \$300 of pilot funding to get your project started
- Collaborative grants of up to \$10,000 are available for larger projects



# Join us at Office Hours!

- Comprehensive online documentation and training resources
- Technical support from a team of scientists, bioinformaticians, and engineers
- **Office Hours on at 10:00 am ET Tuesday and 2:00 pm ET Thursday office hours.**
- **All are welcome!**



# CGC Platform Components

- Key features of running an analysis on CGC
- Several ways to import data to the CGC

# What's the research question you want to answer?



# Prototypical User Flow

Create a Project

Select/access data

Select/create tools

Create and run analysis

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

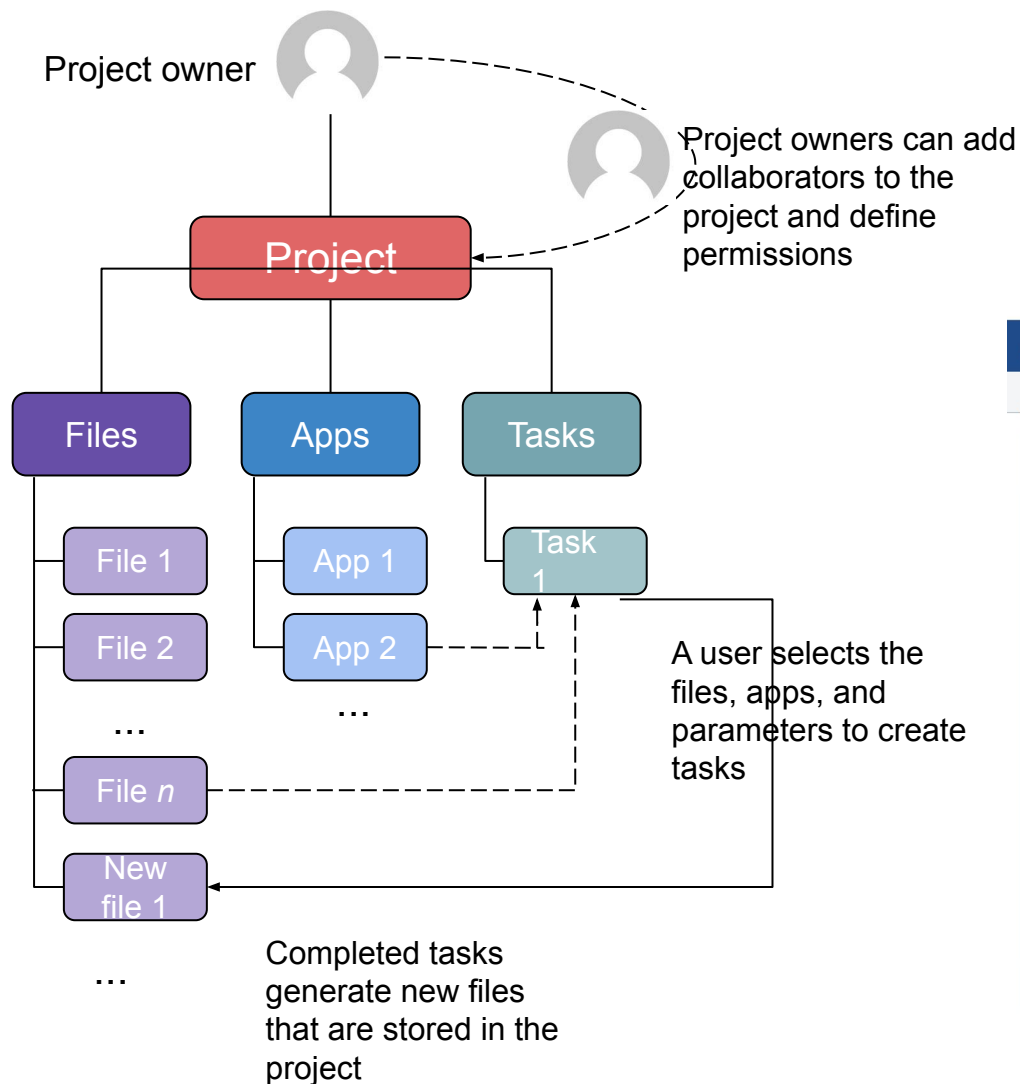
Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

# Projects organize files, methods, and results



Also known as *workspaces* or *sandboxes*  
Easily manage collaborators and permissions

Projects

Data

Public Apps

Public projects

Automations

Developer

Staff

manisha\_ray

Dashboard

Files

Apps

Tasks

Georgetown\_example

Interactive Analysis

Settings

Notes

Description

Welcome to your new project!

Projects are the core building blocks of the CGC Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:

- Start [exploring public datasets](#) straight away
- [Install your tools on the CGC](#) and create workflows
- [Upload your own private data](#) and analyze it along with public datasets
- [Collaborate securely](#) with other researchers

Please record the details of your project here, such as its aims, experimental context, and any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the CGC are logged on the task page. This notepad is just for your own notes.

You can also [use markdown](#) here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the [Knowledge Center](#)

The Seven Bridges CGC Team

Add description

Members

Email notifications

manisha\_ray OWNER

Write, Copy, Execute, Admin

anaDsbg

Write, Copy, Execute

dalibor\_veljkovic

Write, Copy, Execute

Manage members

Analyses

Search

Tasks

Data Cruncher

COMPLETED SBG Decompressor run - 10-24-19 17:35:07

Submitted by dalibor\_veljkovic · Oct. 24, 2019 13:35

BATCH 7 RSEM with STAR Workflow 1.3.1 run - 10-23-19 17:58:39

Submitted by manisha\_ray · Oct. 23, 2019 13:59

COMPLETED RSEM with STAR Workflow 1.3.1 run - 10-23-19 16:36:16

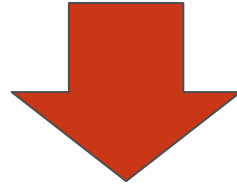
Submitted by manisha\_ray · Oct. 23, 2019 12:36

# Collaborate and share results quickly and easily

The screenshot displays the Cancer Genomics Cloud interface. At the top, there is a navigation bar with a logo and menu items: Projects, Data, Public Apps, Public projects, Developer, and Staff. The user's name 'jack\_digi' is visible in the top right. Below the navigation bar, there is a breadcrumb trail: Dashboard > Files > Apps > Tasks > **CONTROLLED** International Lymphoma Epidemiology Co... > Interactive Analysis > Settings > Notes. The main content area is titled 'Files' and includes a search bar and filter options: Type: All, Sample ID: All, Task ID: All, Tags: All, and a 'Clear filters' button. A table of files is displayed with the following columns: Name, Experimental strategy, Platform, and Type. The table contains several rows of files, including folders and individual files with their respective types. A 'Refresh' button is located at the bottom left of the table, and a pagination indicator 'Showing 1-100 of 207' is at the bottom right. The footer includes links for Forum, Terms, Privacy, and Data Use, along with the copyright notice '© 2019 Seven Bridges Genomics' and a help icon.

| Name  | Experimental strategy | Platform | Type          |
|---|-----------------------|----------|---------------|
| wgs_evaluation_regions.hg38.interval_list<br><small>REFERENCE FILES</small> | -                     | -        | INTERVAL_LIST |
| organized_folder  | -                     | -        | -             |
| hg38_even_handcurated_20k_intervals<br><small>REFERENCE FILES</small>       | -                     | -        | INTERVALS     |
| hapmap_3.3.hg38.vcf.gz.tbi<br><small>REFERENCE FILES</small>                | -                     | -        | TBI           |
| hapmap_3.3.hg38.vcf.gz<br><small>REFERENCE FILES</small>                    | -                     | -        | VCF.GZ        |
| ashkenazim.vcf.gz.tbi   | -                     | -        | TBI           |
| ashkenazim.vcf.bgz  | -                     | -        | BGZ           |

# Prototypical User Flow



Create a Project

Select/access data

Select/create tools

Create and run analysis

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

# Access and search huge public datasets on the CGC

- Web and API based query tools
- Ability to search multiple datasets together



# Access and search huge public datasets on the CGC

The screenshot displays the Cancer Genomics Cloud (CGC) interface. At the top, there is a navigation bar with a logo on the left and menu items: 'Projects', 'Data', 'Public Apps', 'Public projects', and 'Developer'. On the right side of the navigation bar, there is a notification bell icon and the user name 'zworman'. Below the navigation bar, the main content area is titled 'New query' and includes a search bar labeled 'Search by ID'. The primary section is 'Featured datasets', which is expanded to show a list of seven datasets. Each dataset entry includes a checkbox, the dataset name, its version and data model (DM) information, and a 'Details' link.

| Dataset Name  | Version     | DM           | Selected                            | Details |
|---------------|-------------|--------------|-------------------------------------|---------|
| TCGA GRCh38   | GDC v30     | DM 2021.12.1 | <input checked="" type="checkbox"/> | Details |
| TARGET GRCh38 | GDC v30     | DM 2021.12.1 | <input type="checkbox"/>            | Details |
| CPTAC         | DM 2020.8.0 |              | <input type="checkbox"/>            | Details |
| CPTAC-3       | GDC v30     | DM 2021.12.1 | <input type="checkbox"/>            | Details |
| TCIA          | DM 2020.8.0 |              | <input type="checkbox"/>            | Details |
| ICGC          | DM 2021.4.1 |              | <input type="checkbox"/>            | Details |
| FM            | GDC v28     | DM 2021.3.1  | <input type="checkbox"/>            | Details |

# Data Browser feature for building queries and cohorts

TCGA GRCh38 **New query** Edited Create new query Queries ▾ Search by ID Copy files to project

**Case**  
Primary site  
Breast

**Sample**  
Sample type  
Primary Tumor  
Solid Tissue Normal

**File**  
Access level  
Open  
Data type  
Gene Expression Quan...

**Analysis**  
Workflow type  
HTSeq - Counts  
HTSeq - FPKM-UQ  
Filter by  
Properties and values  
Analysis identifier  
Created datetime  
Submitter ID  
Updated datetime  
Workflow link  
Workflow version

**Case** 1,093 ▾ **Sample** 1,211 ▾ **File** 2,432 ▾ **Analysis** 2,432 ▾

Export ▾ Details Analytics

# Conveniently bring in your own data

## Data Tools

Manage your data using any of the following tools to suit your various requirements

### Seven Bridges Command Line Interface (SB CLI)

Upload your data using our fast and secure upload client, taking advantage of parallelization where possible.  
[Learn more](#)

[Download](#)

### Seven Bridges File System (SBFS) BETA

Mount your projects and use files locally or download the executable.  
[Learn more](#)

```
curl https://igor.sbgenomics.com/downloads/sbfs/install.sh -sSf | sudo sh
```

[Download](#)

### Upload files via the API

Upload files using the Seven Bridges Python library.  
[Learn more](#)

```
files = [\n    '/foo/bar/baz.bam'\n    '/foo/bar/qux.fastq'\n]\nfor file in files:\n    api.files.upload(project='my-project', path=file)
```

New folder + Add files

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

Size

-

# Conveniently bring in your own data



**Drag & drop files from your computer or**

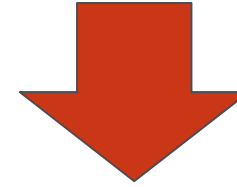
[Browse files](#)

This upload method is primarily intended for small-scale uploads.  
To upload a [larger volume of files](#), please use our [Data Tools](#). Learn more about [uploading from your computer](#).

New folder + Add files ▼ ...

| Size |                                     |
|------|-------------------------------------|
|      | 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         |

# Prototypical User Flow



Create a Project

Select/access data

Select/create tools

Create and run analysis

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

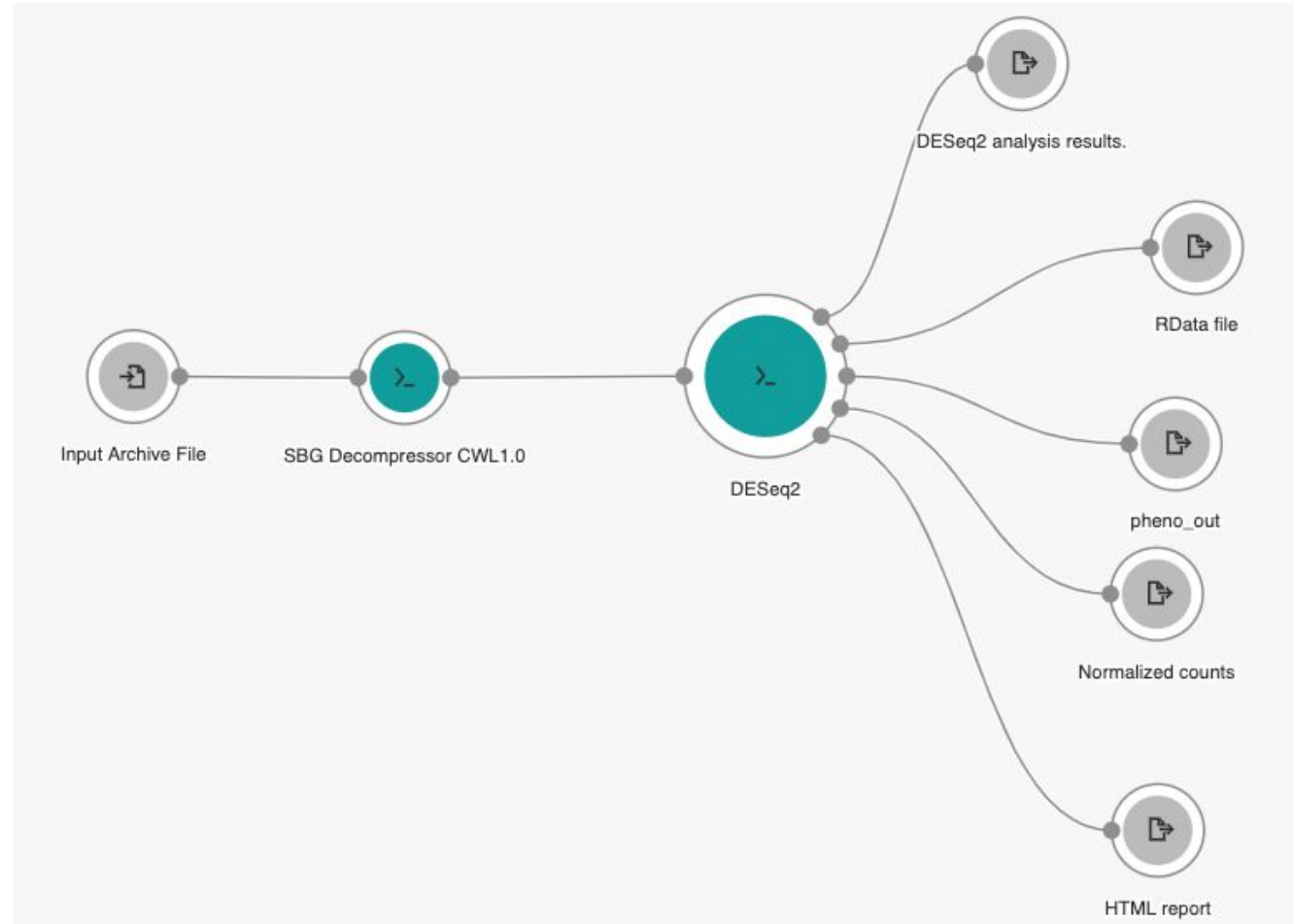
# Find the tools you need in the Public Apps Gallery

- A curated collection of **600+** bioinformatics tools & workflows
  - Optimized for speed & cost in the cloud
  - Fully parameterized & customizable
  - Accessible via the GUI & API

The screenshot displays the 'Public apps' section of a web application. At the top, there is a navigation bar with tabs for 'Data', 'Public Apps', 'Public projects', 'Automations', 'Developer', and 'Staff'. Below this, the 'Public apps' header is visible. A search bar contains the text 'Search workflows and tools'. To the right, a dropdown menu shows 'Category: Differential-Expression ^'. Further right, there are 'Toolkit' and 'Reset search' options. A modal window is open, showing a search for 'Differential-Expression' categories. The modal has a search bar and a 'Clear selected' button. Below the search bar is a grid of categories: Alignment, Analysis, Annotation, Assembly, BED-Processing, CWL1.0, ChIP-seq, Characterization, Converters, Copy Number Variant Calling, Copy-Number-Analysis, DNA, DNA-Methylation, Differential-Expression (highlighted), Enrichment, FASTA-Processing, FASTQ-Processing, Fusions, GATK-4, Genomics, HLA-typing, Imaging, Indexing, Metagenomics, and MiRNA, Microsatellites, Other. Below the modal, several tool cards are visible: 'Ballgown 2.8.4' (Differential-Expression), 'Cufflinks 2.2.1' (Differential-Expression, SAM/BAM-PROCESSING), 'Cuffnorm' (Cufflinks 2.2.1), 'Cuffquant' (Cufflinks 2.2.1), and 'CummeRbundQC' (CummeRbund 2.8.2). Each card includes a brief description of the tool's function.

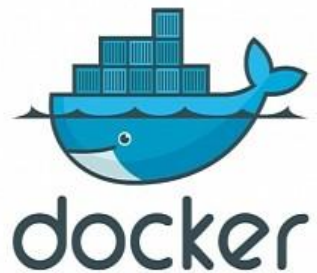
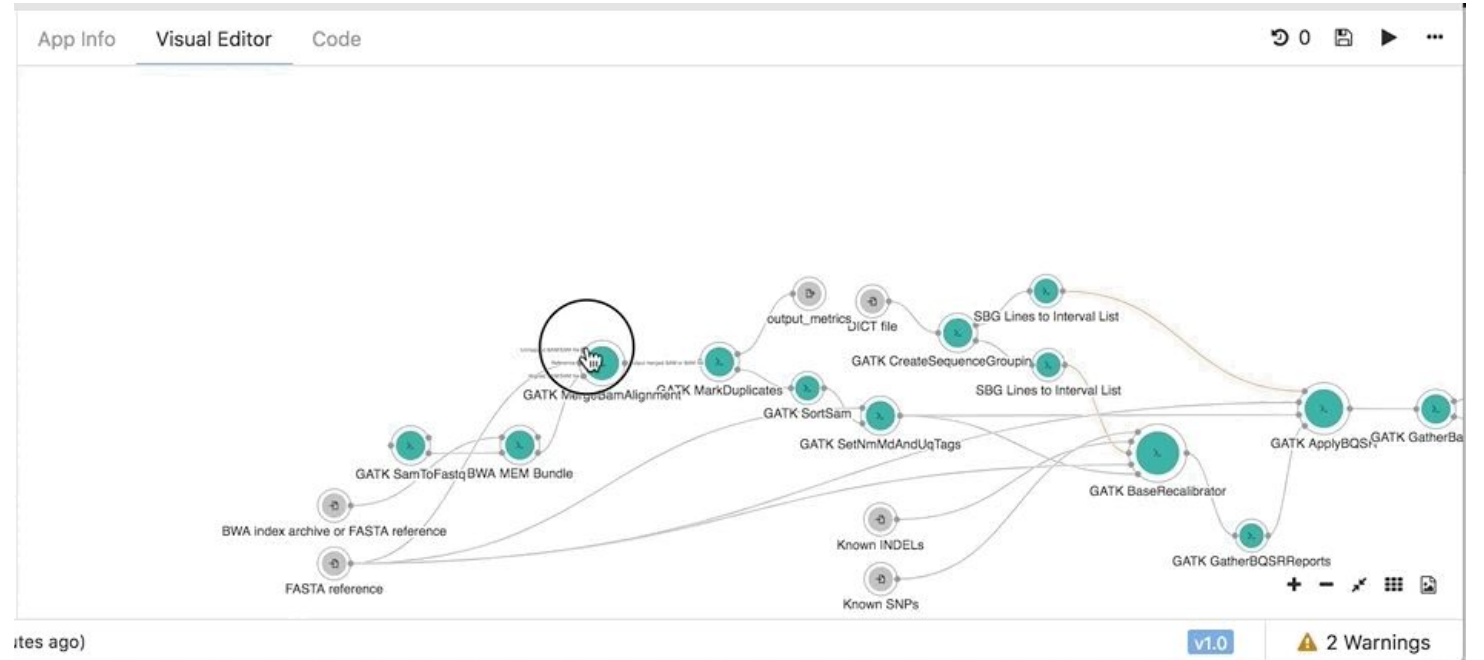
# Use Web Composer to tailor a workflow

- Graphical, easy to use interface



# Bring your pipelines to the platform with Web Composer

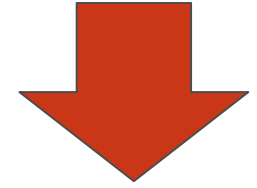
- An intuitive and flexible software development kit for developing and porting custom tools to the platform
- Conformance with community standards to ensure pipeline portability & reproducibility



COMMON  
WORKFLOW  
LANGUAGE

**Rabix**  
[Reproducible Analysis for Bioinformatics]

# Prototypical User Flow



Create a Project

Select/access data

Select/create tools

Create and run analysis

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Tools, workflows, and software packages

- Public Apps Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

# Set up tasks to analyze your data

Cloud icon | Projects | Data | Public Apps | Public projects | Automations | Developer | Staff | manisha\_ray

Dashboard | Files | Apps | **Tasks** | Georgetown\_example | Interactive Analysis | Settings | Notes

**BATCH 7 RSEM with STAR Workflow 1.3.1 run - 10-25-19 04:32:48** [edit]

Last update by manisha\_ray on Oct. 25, 2019 00:32  
App: RSEM with STAR Workflow 1.3.1 - Revision: 4

Task Inputs | Execution Settings

### Inputs

Batching  On

Input file \* [Change selection]

Batch by: File

This will create one task for each selected item.

- G20502.22Rv1.2.bam (1 item)
- G20506.DU\_145.2.bam (1 item)
- G26174.NCI-H660.2.bam (1 item)
- G27214.PC-3.1.bam (1 item)
- G28030.MDA\_PCa\_2b.1.bam (1 item)
- G28033.LNCaP\_clone\_FGC.1.bam (1 item)
- G41666.VCaP.5.bam (1 item)

### App Settings

[Edit parameters] [Show editable]

RSEM Prepare Reference (#RSEM\_Prepare\_Reference)

Bowtie: No value

Bowtie 2: No value

STAR: True

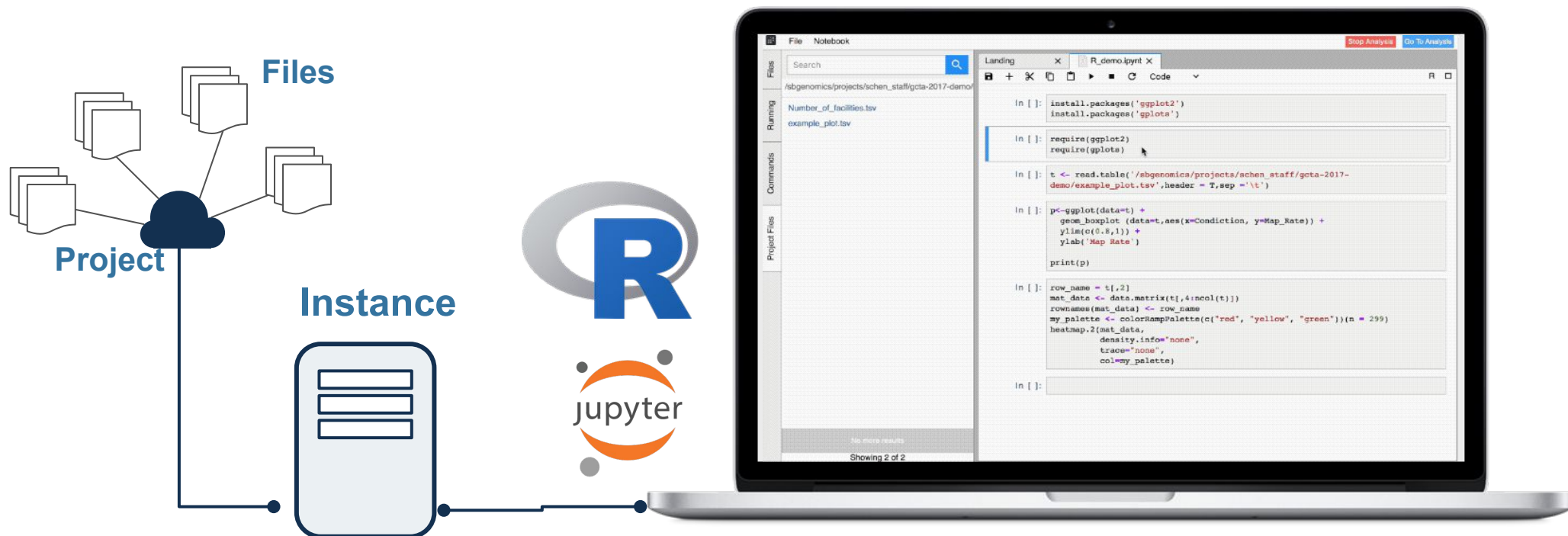
STAR splice junction database overhang: [ ]

### Outputs

|                          |          |
|--------------------------|----------|
| First strand             | No value |
| Genes results            | No value |
| Genome BAM               | No value |
| Isoforms results         | No value |
| RSEM Plot Model PDF File | No value |
| STAR Log Files           | No value |
| STAR splice junctions    | No value |
| Second strand            | No value |
| Transcript BAM           | No value |

# Powerful, collaborative, & reproducible interactive analysis

Users create interactive analysis sessions within a project - all files are available and over 50 instances can be used (*c3.xlarge* to *x1.32xlarge* on AWS)



# Quantify and visualize expression differences in RStudio

Data Public Apps Public projects Developer Staff manisha\_ray

ps Tasks isomir test analysis Copy pro

## Explore genomics

Understand complex genomic data

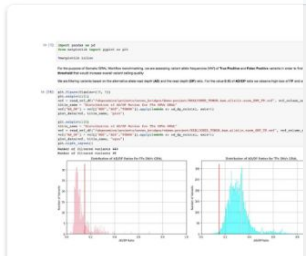


### Genome Browser

Visualize alignments, SNV/Indels, annotation tracks, check coverage and mismatch, assess alignments and variants

0 files

Open



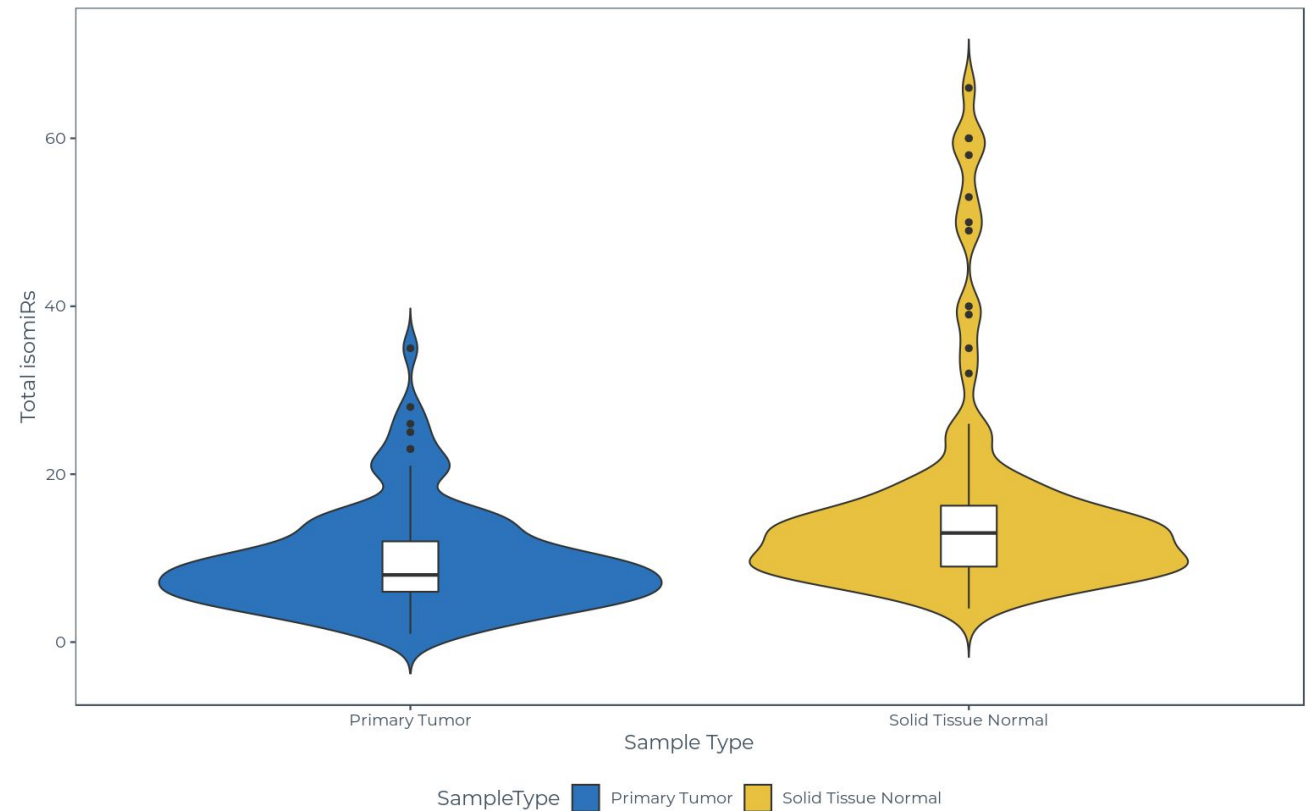
### Data Cruncher

Analyze and explore data using JupyterLab or RStudio

Open

```
R File Edit Code
logic.R app.R
1 # Copyright (C) 201
2 #
3 # This document is
4 # It is considered
5 #
6 # This document may
7 # in whole or in pa
8 # Seven Bridges Ge
9
10 libpath <- "/sbgenc
11 .libPaths(libpath)
12
13 library("shiny")
14 library("shinydasht
15
16 library("sevenbridg
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

IsomiRs by sample type - hsa-let-7c-3p



Data Use



# Using the CGC for RNA sequencing analysis



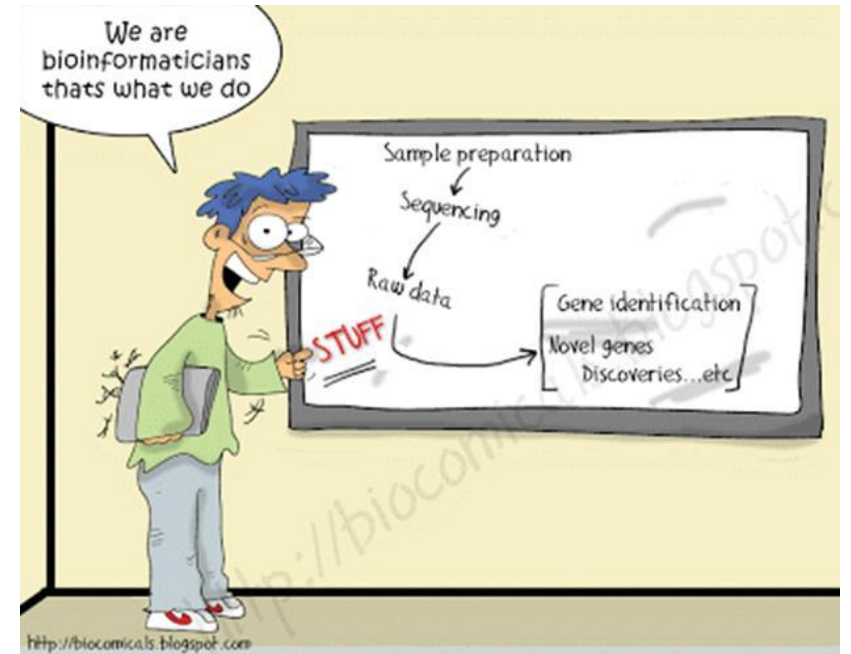
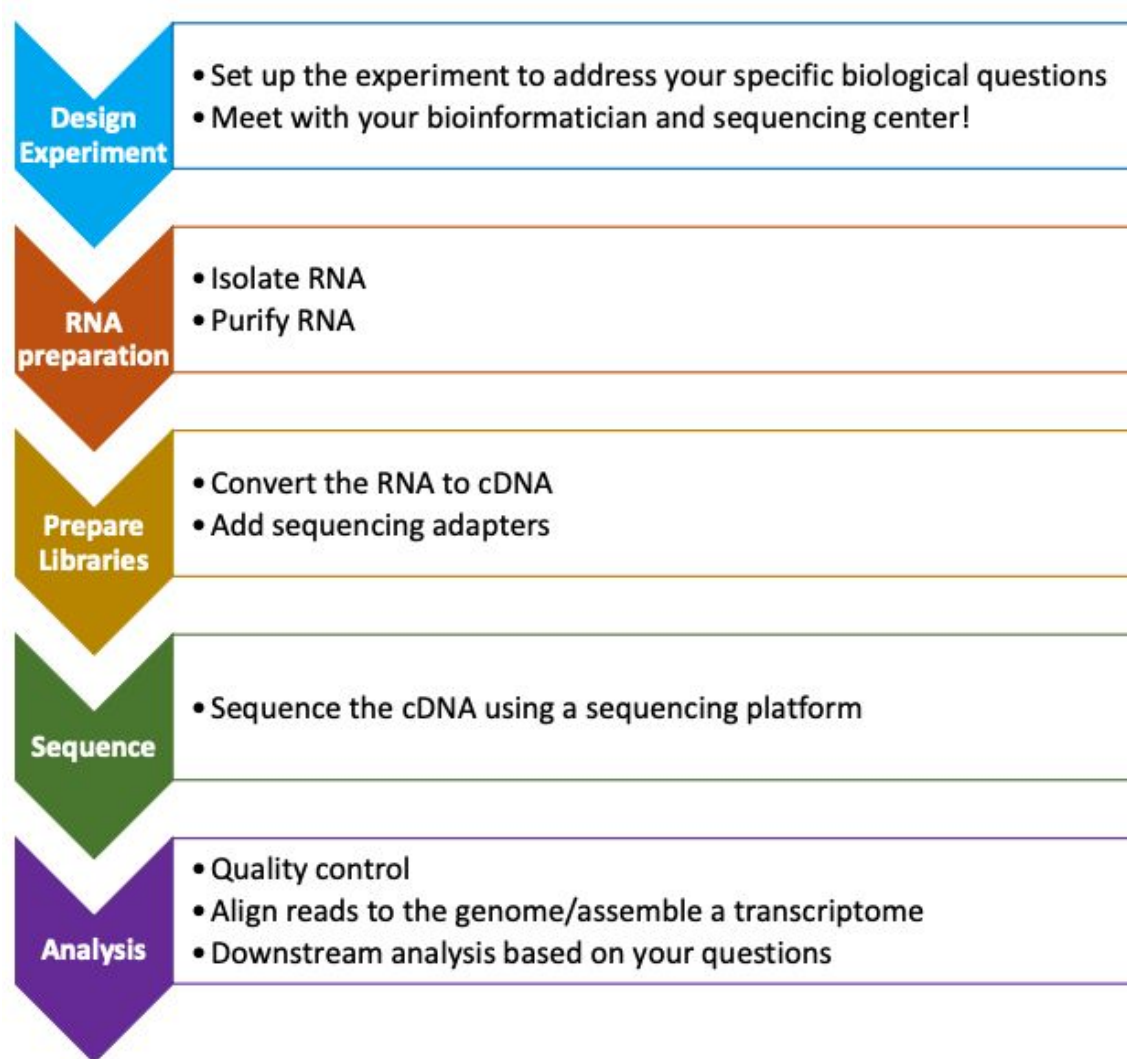
# What is RNAseq?

- A high-throughput sequencing-based method to quantify gene expression in a sample
- All of the RNA transcripts in a sample (transcriptome) are sequenced
  - Sequencing data is aligned to a reference genome or transcriptome
  - Allows for quantification of gene expression and differential expression analyses and characterization of alternative splicing
  - *de novo* transcriptome assembly - no genome sequence necessary!

# Why do RNA sequencing?

- Unbiased - collect all of the RNA present in the sample
  - Allows for discovery of new gene targets and pathways
  - In contrast to **targeted** methods where you must choose the genes of interest in advance
- Dynamic - collect all the genes that are made at the time of sample collection
  - Monitor changes over time (ex: disease progression)
  - Monitor changes in response to treatment (ex: drug effects)

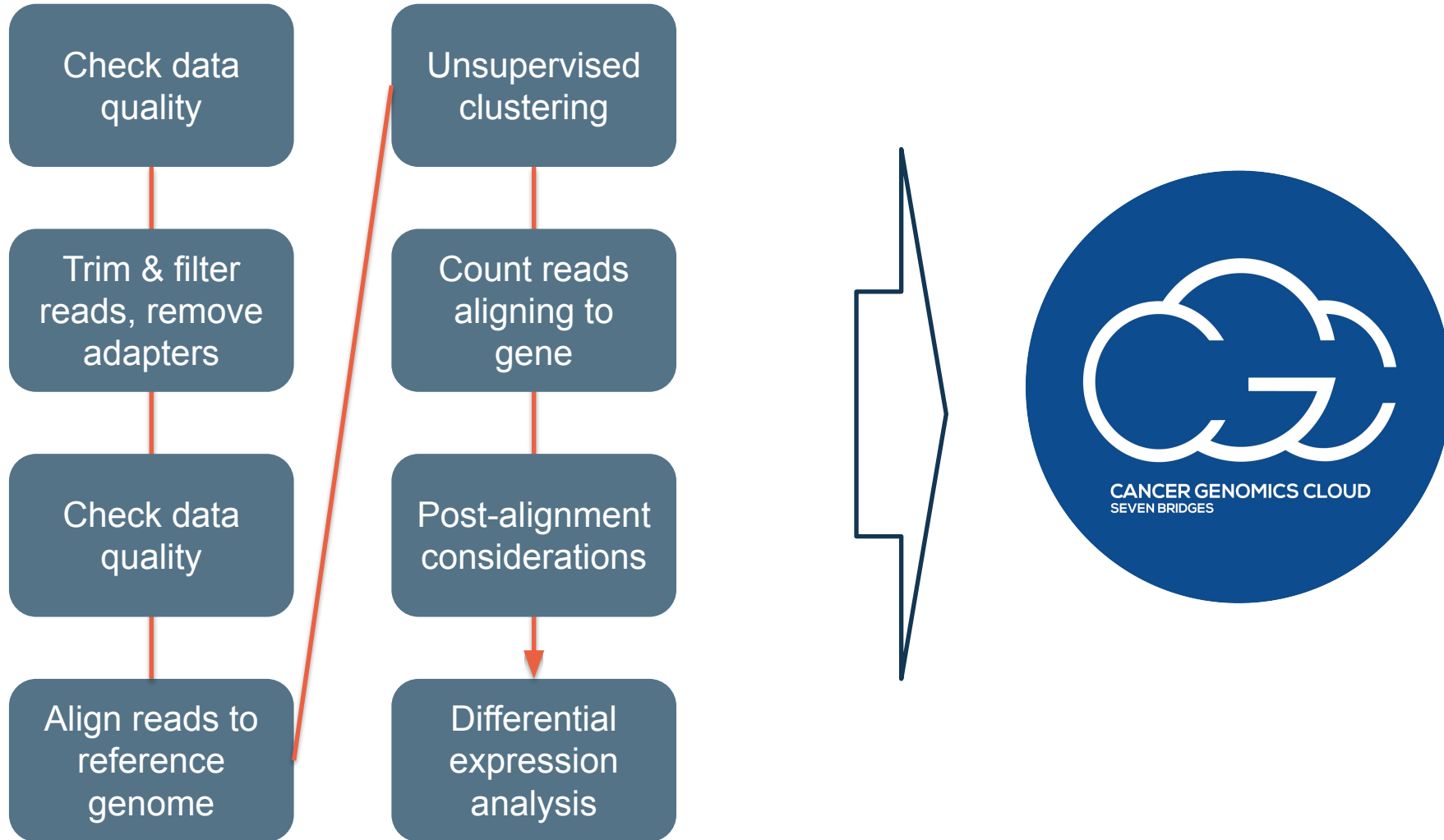
# RNAseq workflow overview



We are here!

Slide adapted from Dr. Nadia Attalah, Purdue University

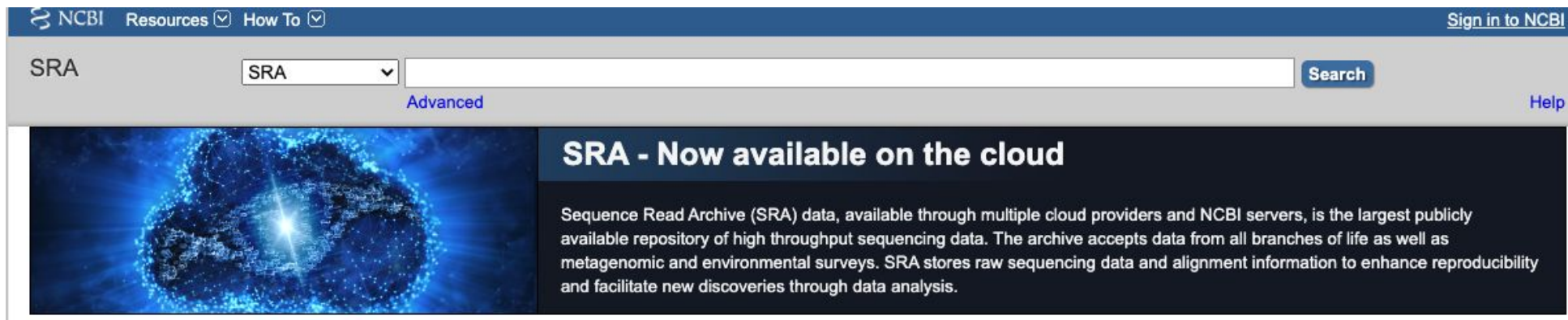
# Standard Differential Expression Analysis



# What is SRA?

- SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.
- Search by author, pubmed ID, etc.
- Download respective accession list and numbers, sequence data files using SRA Toolkit, and associated metadata associated with SRA data

<https://www.ncbi.nlm.nih.gov/sra>



NCBI Resources How To Sign in to NCBI

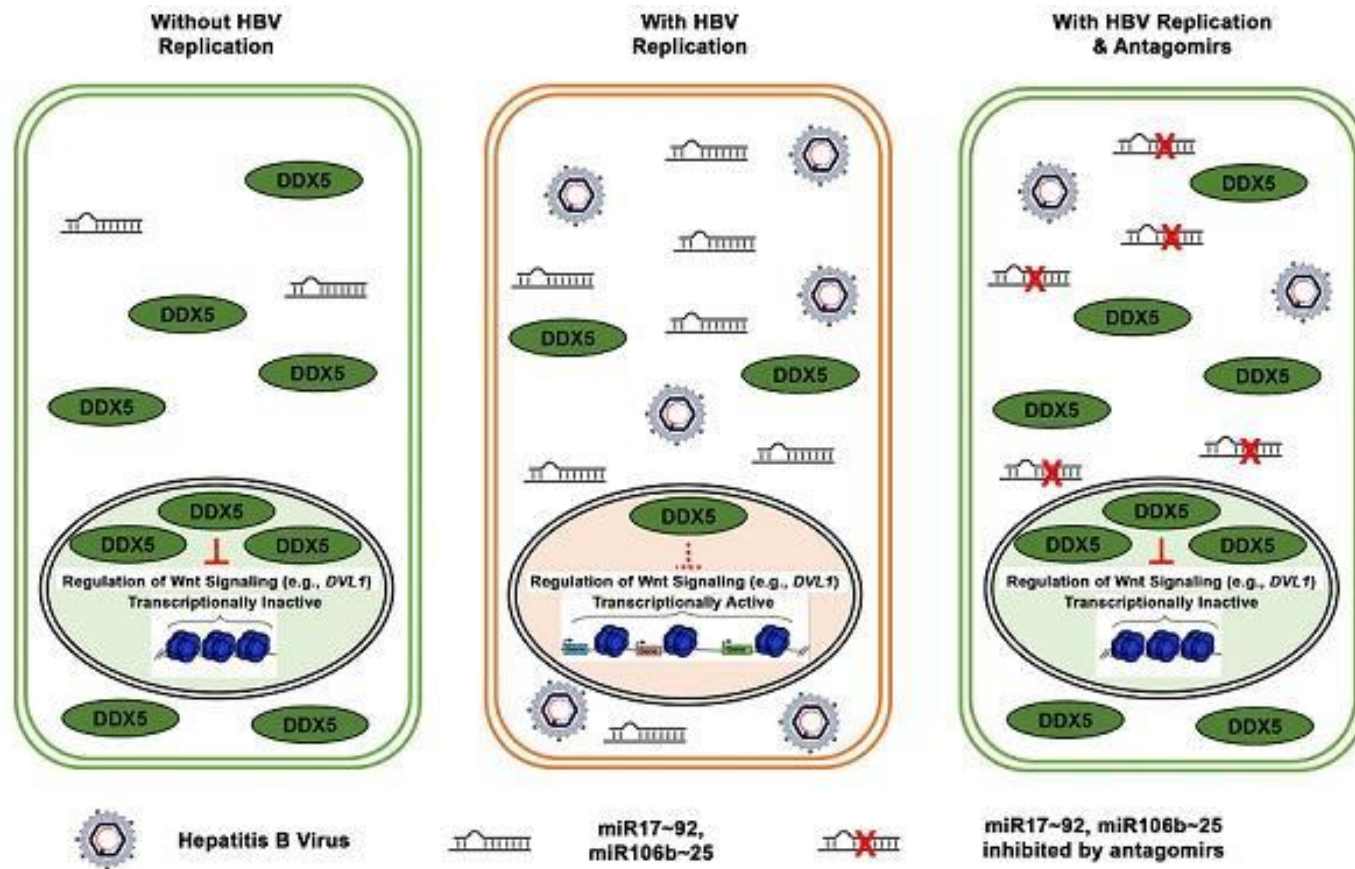
SRA SRA Search

Advanced Help

### SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

# Overview of data we will use for bulk RNA-seq analysis



Mani SKK, et al. Restoration of RNA helicase DDX5 suppresses hepatitis B virus (HBV) biosynthesis and Wnt signaling in HBV-related hepatocellular carcinoma. *Theranostics* 2020; 10(24):10957-10972. doi:10.7150/thno.49629. <https://www.thno.org/v10p10957.htm>

## Data availability

All sequencing data are available through the NCBI Gene Expression Omnibus (GEO) database (accession number **GSE131257**).

# Search SRA with GEO number!

NCBI SRA Run Selector

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Log in to NIH

Accession: GSE131257 Search

**Filters List**

- 1  Assay Type
- 2  AvgSpotLen
- 3  Bases
- 4  Bytes
- 5  Cell\_line
- 6  Genotype
- 7  genotype/variation
- 8  induced\_hbv\_replication
- 9  LibrarySelection
- 10  LibrarySource
- 11  source\_name

**Common Fields**

|                  |                                  |
|------------------|----------------------------------|
| DATASTORE region | gs.US, ncbi.public, s3.us-east-1 |
| Instrument       | Illumina HiSeq 2500              |
| LibraryLayout    | PAIRED                           |
| Organism         | Homo sapiens                     |
| Passage          | 20-23                            |
| Platform         | ILLUMINA                         |
| ReleaseDate      | 2020-10-07                       |
| SRA Study        | <a href="#">SRP198483</a>        |
| tissue           | Liver                            |

downloads file "SraRunTable.txt - Metadata"

**Select**

|          | Runs | Bytes    | Bases    | Download                               | Cloud Data Delivery | Computing |
|----------|------|----------|----------|--|---------------------|-----------|
| Total    | 19   | 78.31 Gb | 242.91 G | Metadata or Accession List             |                     |           |
| Selected | 0    | 0        | 0        | Metadata or Accession List or JWT Cart | Deliver Data        | Galaxy    |

downloads file "SRR\_Acc\_list.txt"

# On the CGC, upload metadata, tell the tool the SRA accession...

The screenshot displays the CGC interface for a task named "SRA Download and Set Metadata run - 02-18-22 17:50:52". The top navigation bar includes "Projects", "Data", "Public Apps", "Public projects", and "Developer". The user "zworman" is logged in. The task is currently in the "Tasks" view. Below the task title, there are buttons for "Get support", "View stats & logs", and "Edit and rerun".

Execution details: Executed on Feb. 18, 2022 12:53 by zworman. Spot Instances: On. Memoization (WorkReuse): On. Price: \$3.37. Duration: 3 hours, 59 minutes.

App: SRA Download and Set Metadata - Revision: 0

**Inputs**

- SRA metadata file: metadata\_rnaseq-purdue.txt

**App Settings**

- SBG Prepare Metadata for SRA download** (#sbg\_prepare\_metadata\_for\_sra\_download\_1)
  - Accession (SRA accession): SRP198483
- SRA fasterq-dump (adjusted)** (#sra\_fasterq\_dump\_adjusted\_v2\_10\_7)
  - Number of threads: 4
  - Split 3: False
  - Split files: True

**Output Settings**

- Metadata file
- Out reads

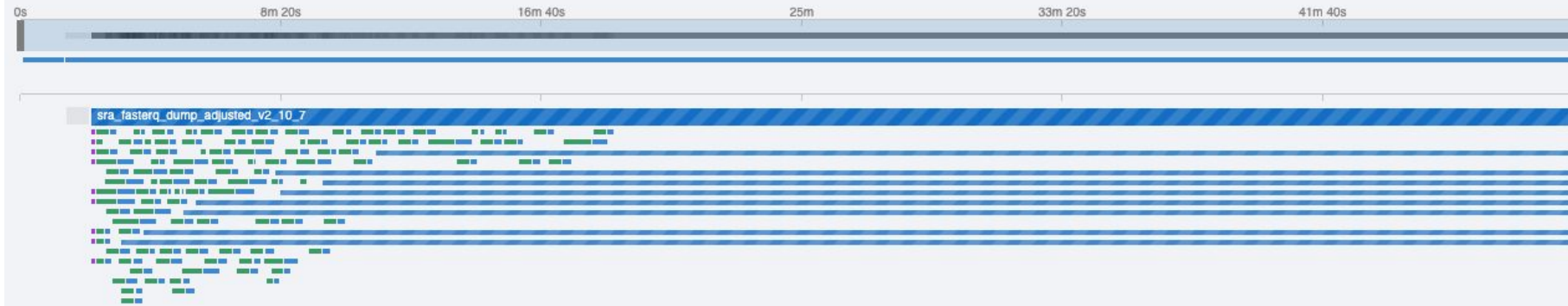
# RNAseq data transferred from SRA to CGC

- Total time to transfer 3 hours, 59 minute, \$3.37

**RUNNING** Tasks / SRA Download and Set Metadata run - 02-18-22 13:51:55 / Stats

[Instance metrics](#) [View task logs](#)

Search apps 



## Quick Details

N/A

|             |     |
|-------------|-----|
| Start Time: | N/A |
| End Time:   | N/A |
| Duration:   | N/A |
| Instances:  | N/A |

## Pinned Details

|  |   |
|--|---|
| <b>sra_fasterq_dump_adjusted_v2_10_7</b> |  |
| Start Time:                              | 2m 15s [08:54:15]   |
| End Time:                                | 49m 58s [09:41:58]  |
| Duration:                                | 47m 43s   |
| Instances:                               | c4.8xlarge (1024GB) [spot]  |
| Status:                                  | <b>RUNNING</b>  |

[View Logs](#)

# RNAseq data transferred from SRA to CGC

Dashboard **Files** Apps Tasks **Purdue Lecture Test 1** Interactive Analysis Settings Notes

Files > bulk-rnaseq-samples New folder + Add files ...

Search Extension: All Tags: All + Clear filters

| <input type="checkbox"/> Name                | Experimental strategy | Extension | Size     | Sample ID   |
|--|-----------------------|-----------|----------|-------------|
| <input type="checkbox"/> SRR12776583_1.fastq | -                     | FASTQ     | 61.7 GiB | SRR12776583 |
| <input type="checkbox"/> SRR12776583_2.fastq | -                     | FASTQ     | 61.7 GiB | SRR12776583 |
| <input type="checkbox"/> SRR12776584_1.fastq | -                     | FASTQ     | 72.2 GiB | SRR12776584 |
| <input type="checkbox"/> SRR12776584_2.fastq | -                     | FASTQ     | 72.2 GiB | SRR12776584 |
| <input type="checkbox"/> SRR12776585_1.fastq | -                     | FASTQ     | 43.5 GiB | SRR12776585 |
| <input type="checkbox"/> SRR12776585_2.fastq | -                     | FASTQ     | 43.5 GiB | SRR12776585 |
| <input type="checkbox"/> SRR12776586_1.fastq | -                     | FASTQ     | 74.4 GiB | SRR12776586 |
| <input type="checkbox"/> SRR12776586_2.fastq | -                     | FASTQ     | 74.4 GiB | SRR12776586 |
| <input type="checkbox"/> SRR9058988_1.fastq  | -                     | FASTQ     | 3.4 GiB  | SRR9058988  |
| <input type="checkbox"/> SRR9058988_2.fastq  | -                     | FASTQ     | 3.4 GiB  | SRR9058988  |

# RNA seq differential expression

Create a Project

Organizational unit  
within platform

Select/access data

**Upload data** from a  
collaborator's project

Select/create tools

Select **differential  
expression** tools from  
Public Apps Gallery

Connect into a  
workflow

Create and run  
analysis

Set up task and run  
workflow

Review outputs:  
**differentially  
expressed genes**

# Summary: Run analyses in the cloud with the CGC

- CGC: platform for analysis, part of CRDC
- Collaborate seamlessly
- Access petabytes of public datasets
- Build workflow from new or existing tools
- Run tasks using the flexibility and scale of the cloud
- Connect to other data repositories
- Analyze data interactively using notebooks in the cloud



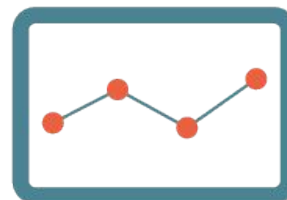
Easy data  
management



Secure  
collaboration &  
managed billing



Flexible & fully  
reproducible  
methods



Optimized  
bioinformatics  
algorithms



Scalable  
computation



Extensible &  
developer  
friendly tools



CANCER GENOMICS CLOUD  
SEVEN BRIDGES

# Let's get started!

Login to the platform at  
[cgc.sbgenomics.com](https://cgc.sbgenomics.com)



CANCER GENOMICS CLOUD  
SEVEN BRIDGES

## Log in



Log in with eRA Commons

[Log in with username and password](#)

New to the CGC? [Create an account](#)

# Q&A and Discussion

