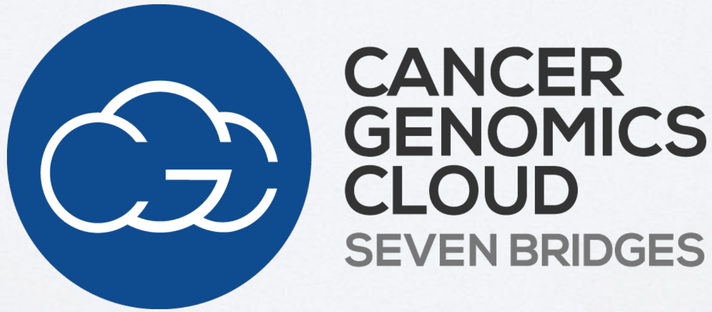
Cancer Genomics Cloud & R: Find, Access, and Analyze Petabyte-Scale Cancer Genomic Data on the Cloud

Nan Xiao
Seven Bridges



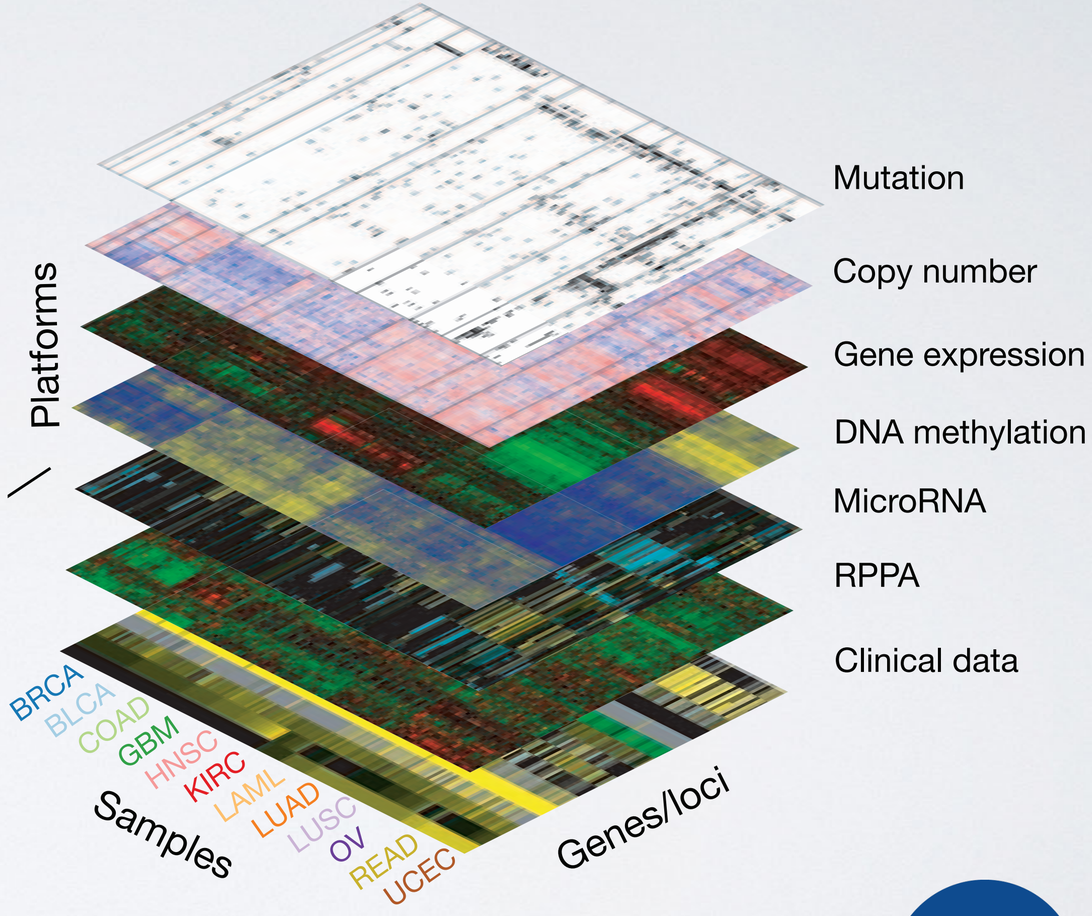
TCGA IS A TREMENDOUS GIFT TO THE CANCER RESEARCH COMMUNITY …

More than 11,000 cases representing 33 cancer types

CANCER GENOMICS CLOUD
SEVEN BRIDGES

UNDERSTANDING TCGA DATASET

- Multiple Samples per Case
- Multiple Analyses per Sample
- Rich metadata: barcodes, UUIDs, XMLs...



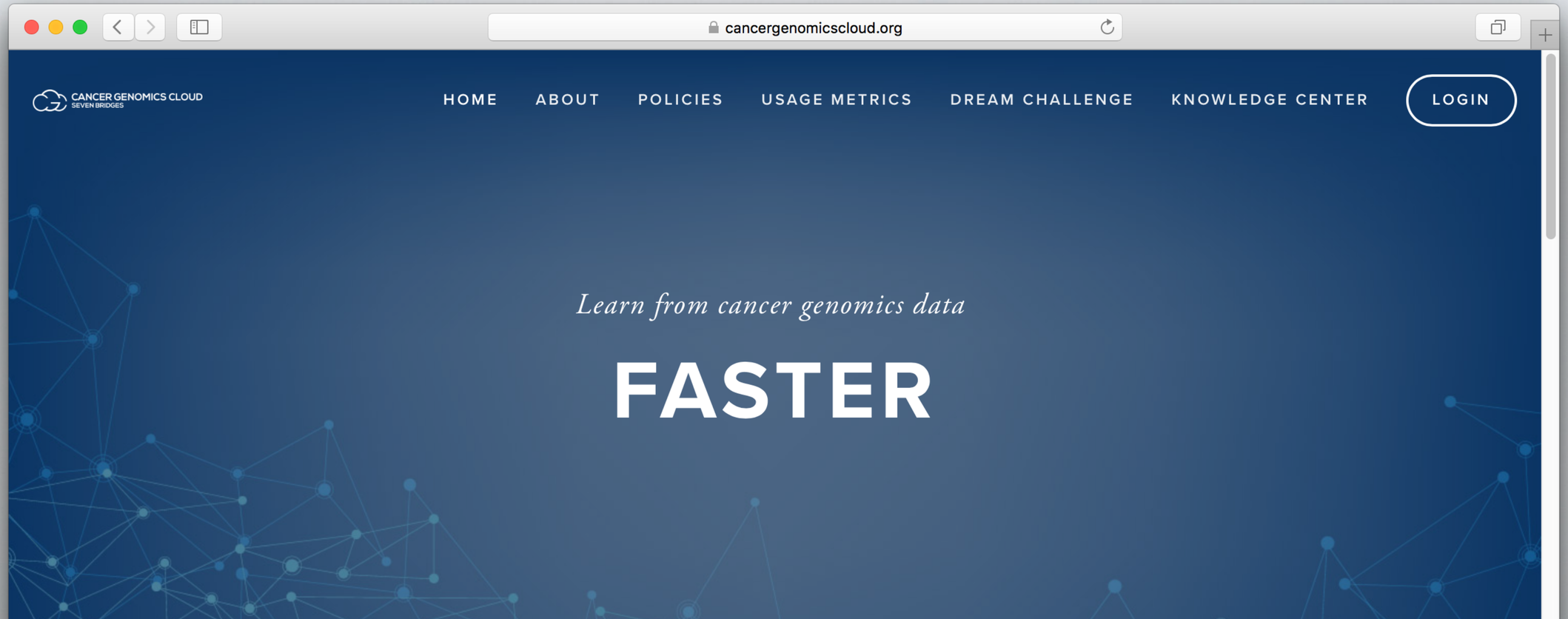
Nature Genetics 45, 1113-1120 (2013)

CANCER GENOMICS CLOUD
SEVEN BRIDGES

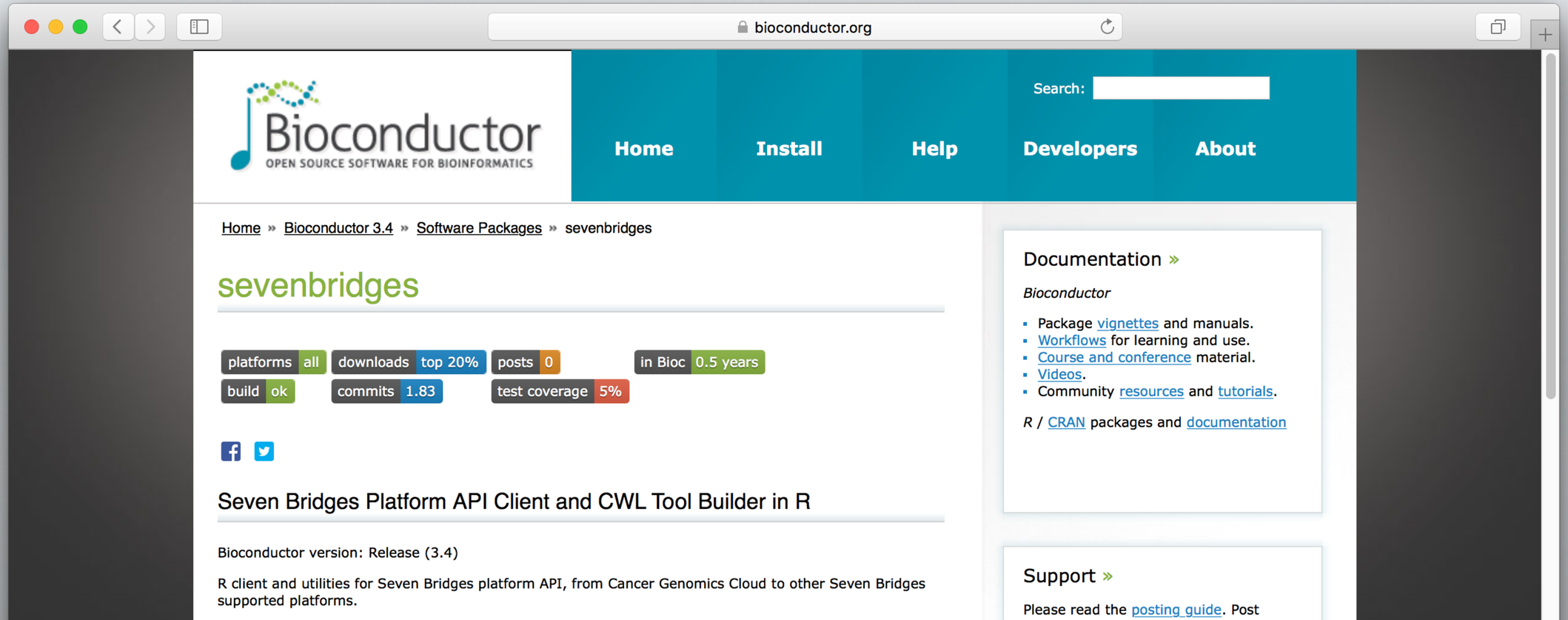
As the amount and diversity of data increases, it becomes more difficult to learn from them.

CANCER GENOMICS CLOUD
SEVEN BRIDGES

The CGC aims to provide a collaborative environment where researchers can take advantage of co-localized public data (like TCGA) and public tools; but also recombine these with their private data and tools.

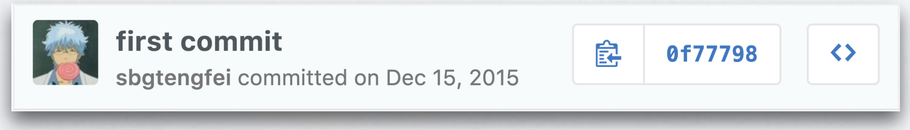


`sevenbridges` is an R/Bioconductor package offering easier programmatic access to CGC.



The `sevenbridges` package offers:

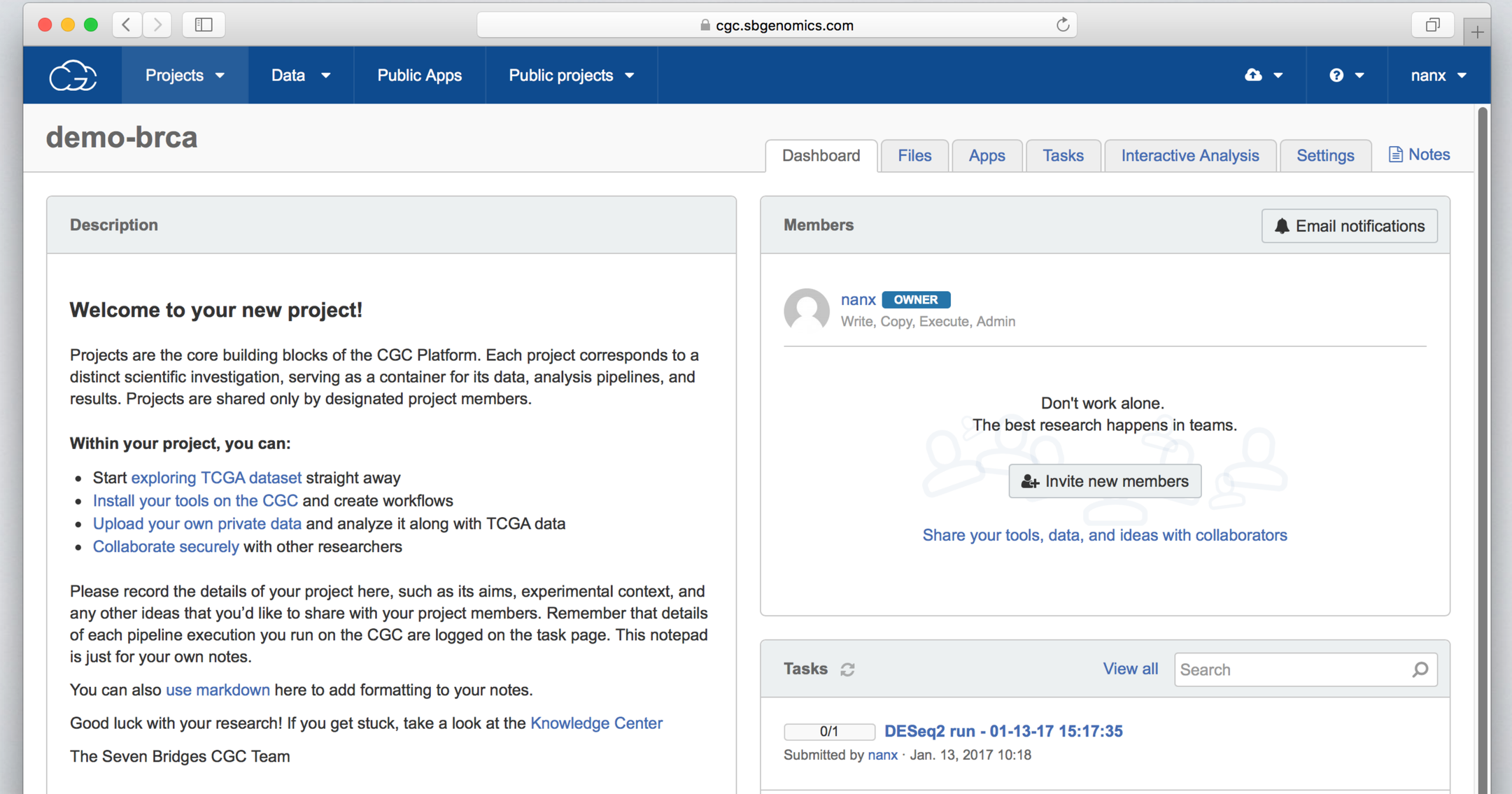
- Complete R API client for CGC/Seven Bridges API
- Common Workflow Language Tool Interface
- Task monitoring / Batch tasks support
- and many helper functions/utilities in R.



2015

2017

PROJECT MANAGEMENT

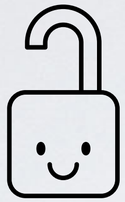


PROJECT MANAGEMENT VIA API

```
library("sevenbridges")

# Create Auth object
a = Auth(platform = "cgc", token = “your_token")

# List project and details
a$project(owner  = "user")
a$project(detail = TRUE)

# Get billing group ID
bid = a$billing()$id

# Create new project
a$project_new("brca_test", bid,
              description = "BRCA Test")

# Add member to project
m = a$project(id = "demo/brca_test")$
      member_add(username = "new_user")
```

FIND DATA ON CGC

MORE THAN ONE PETABYTE OF TCGA DATA AT YOUR FINGERTIPS



Open Data

Information NOT unique to an individual.

- de-identified clinical data
- gene expression data
- copy number alterations
- epigenetic data

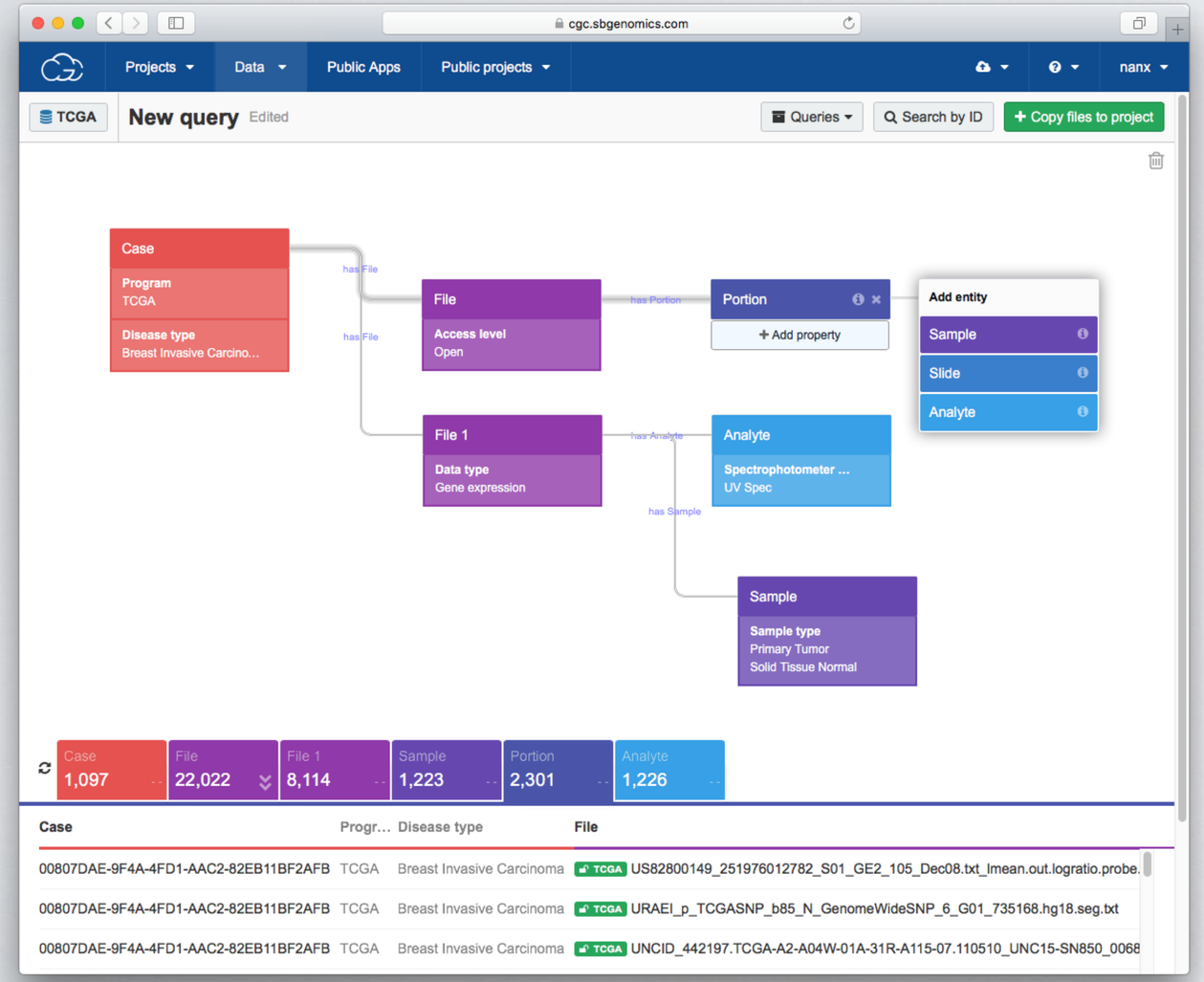
Controlled Data

Information that IS unique to an individual.

- primary sequencing data
- raw & processed SNP6 array data
- raw exon array data
- mutation calls for an individual

CANCER GENOMICS CLOUD
SEVEN BRIDGES

CGC DATA BROWSER



- Query metadata
- Add files to project

CANCER GENOMICS CLOUD
SEVEN BRIDGES

SPARQL QUERY

```
library("SPARQL")
endpoint = "https://opensparql.sbgenomics.com/blazegraph/namespace/tcga_metadata_kb/sparql"
query = "
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix tcga: <https://www.sbgenomics.com/ontologies/2014/11/tcga#>

select distinct ?case ?sample ?file_name ?path ?xs_label ?subtype_label
where
{
 ?case a tcga:Case .
 ?case tcga:hasDiseaseType ?disease_type .
 ?disease_type rdfs:label 'Lung Adenocarcinoma' .

 ...

}
"
qd = SPARQL(endpoint, query)
df = qd$results
head(df)
```

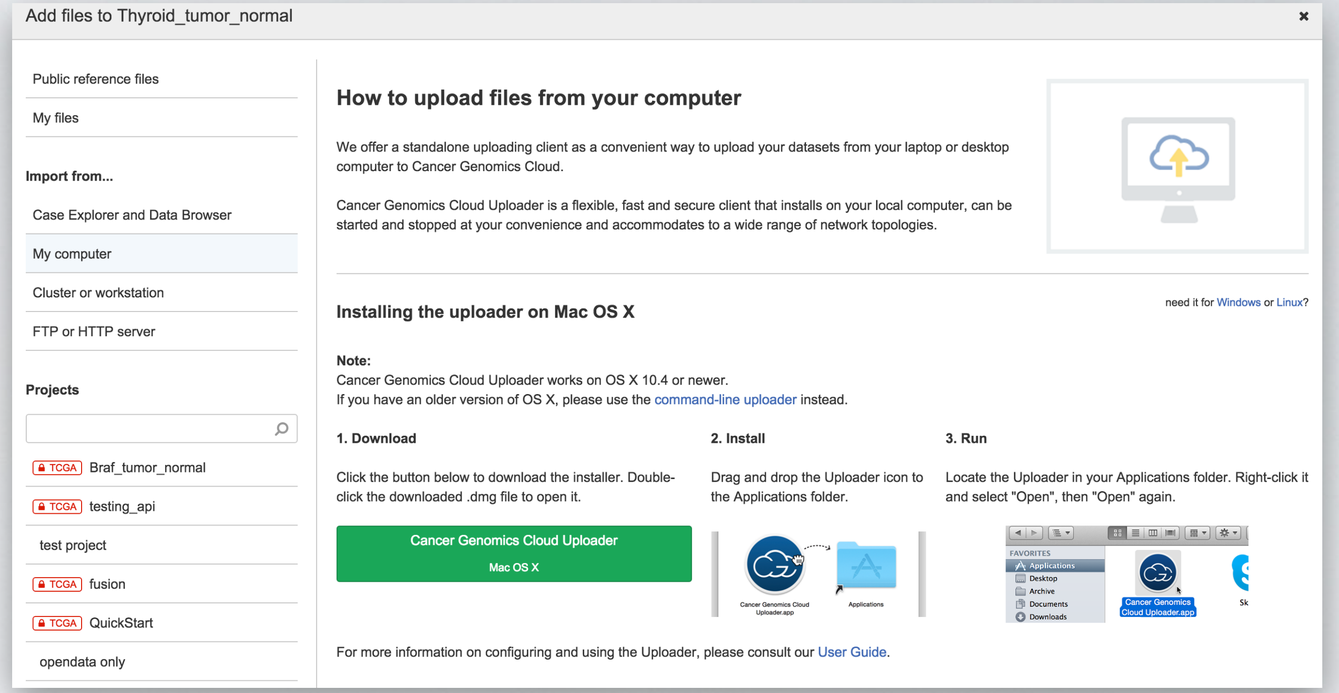
DATASETS API

```
term = list(
    "entity" = "cases",
    "hasSample" = list(
        "hasSampleType" = "Primary Tumor",
        "hasPortion" = list(
            "hasPortionNumber" = 11
            )
        ),
    "hasNewTumorEvent" = list(
        "hasNewTumorAnatomicSite" = c("Liver", "Pancreas"),
        "hasNewTumorEventType" = list(
            "filter" = list(
                "contains" = "Recurrence"
            )
        )
    )
)

a$api(path = "query", body = term, method = "POST")
```

ADD & ANNOTATE DATA

UPLOAD YOUR DATA

- CGC Uploader (GUI)
- Command Line Uploader
- FTP / HTTP / S3 Volume
- API



CANCER GENOMICS CLOUD
SEVEN BRIDGES

UPLOAD A FILE VIA API

```
myfile = "file_path.fastq"
p = a$project(id = "demo/tcga-demo")

# Load `.meta` for the file by default
p$upload(myfile, overwrite = TRUE)

# Pass metadata manually
p$upload(myfile, overwrite = TRUE,
         metadata = list(library_id = "test_id",
                         platform   = "Illumina x11"))

# Rename file
p$upload(myfile, overwrite = TRUE,
         metadata = list(library_id = "new_id")
         name = "sample_new_name.fastq")
```

UPLOAD FILES / FOLDERS

```r
# Upload a folder

dirpath = "path_to_dir"
list.files(dirpath)
p$upload(dirpath, overwrite = TRUE)


# Upload a list of files

dirpath = "path_to_dir"
myfiles = list.files(dirpath, recursive = TRUE, full.names = TRUE)
p$upload(myfiles, overwrite = TRUE)
```

ANNOTATE FILES VIA API

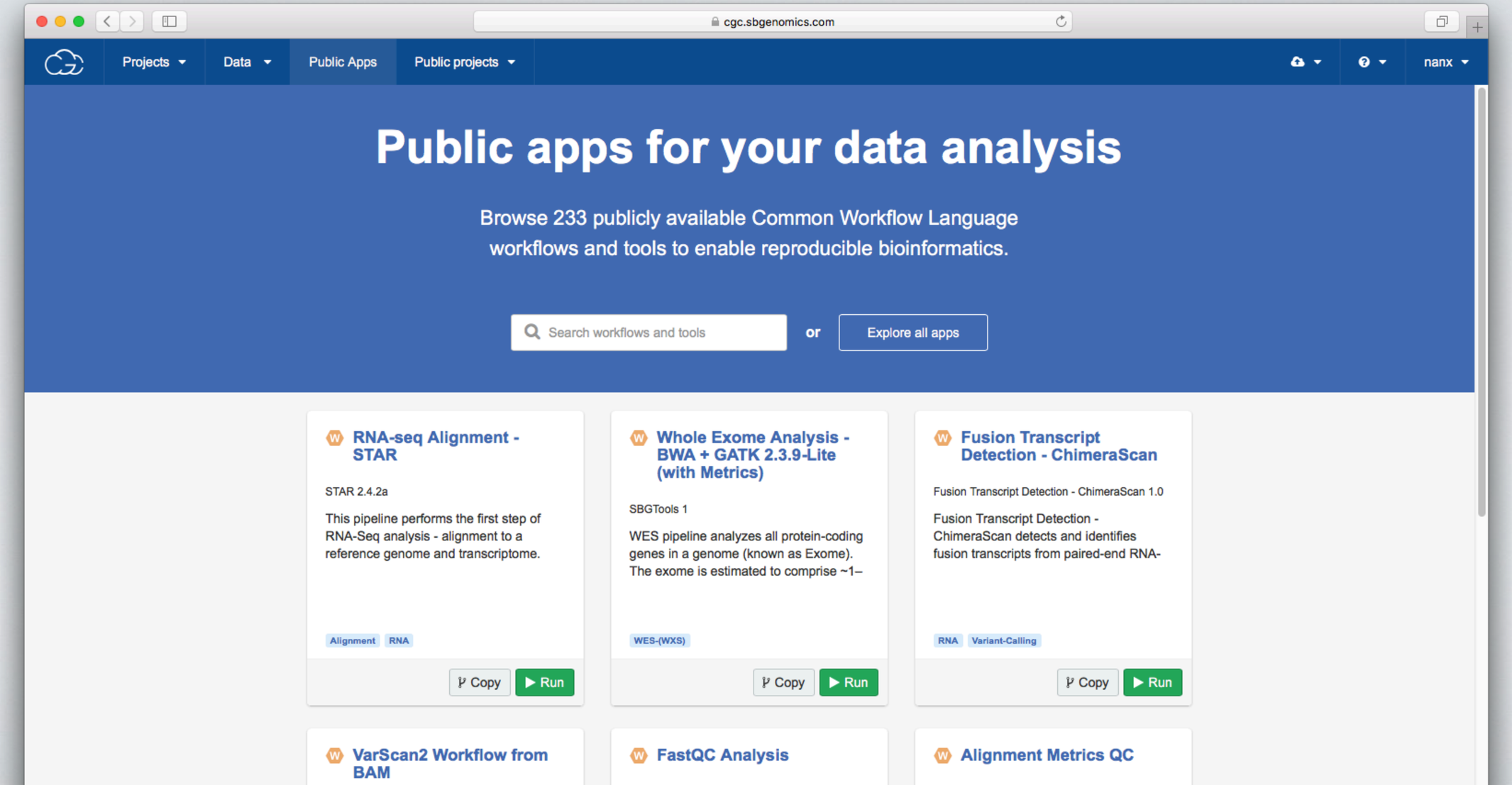
```
# Locate a sample BAM file in project
p  = a$project(id = "demo/tcga-demo")
fl = p$file("sample.bam", exact = TRUE)

# Show tags for single file
fl$tag()

# Add new tags
fl$add_tag("new year new tag")

# Set tags to overwrite existing
x = list("this", "is", 2017)
fl$set_tag(x)

# Set metadata
fl$meta()
fl$set_meta()
```

UPLOAD MANIFEST FILE (DEFINES METADATA OF FILES)

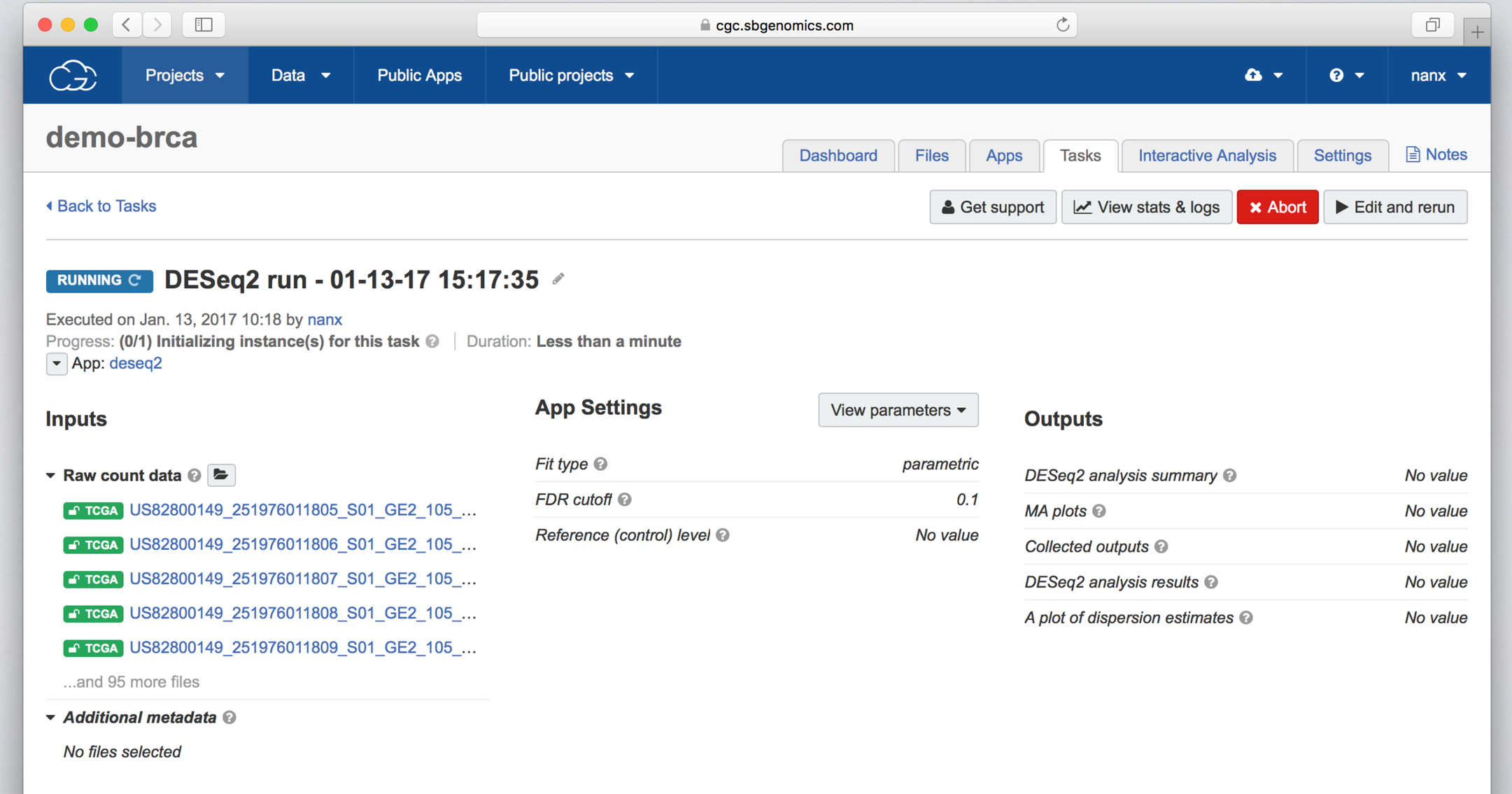
```
# attach all metadata except "bad_field" and "sample_id"

p$upload(manifest_file = "~/manifest.csv",
         overwrite     = TRUE,
         subset        = score < 0.5,
         select        = -c(bad_field, sample_id))
```

ANALYZE THE DATA

Run an analysis immediately, with ~230 tools and workflows on the CGC today.



RUNNING ANALYSIS TASKS



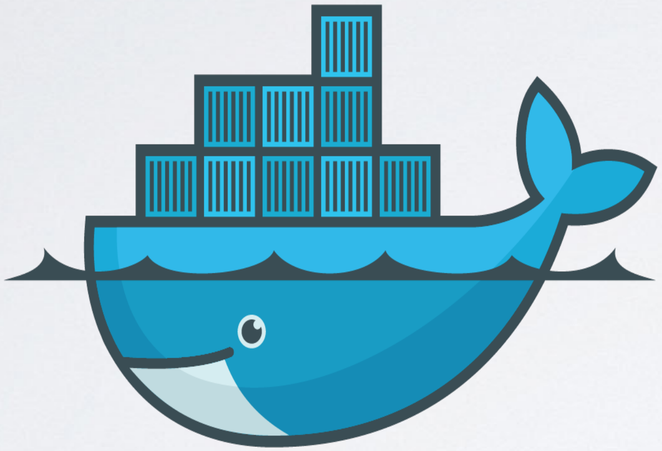
COPY APPS & RUN TASKS

```
# List public apps
a$app(visibility = "public")

# Copy an app to project
aid = a$public_app()[[1]]$id
a$copy_app(aid, project = pid, name = "new_app_name")

# Add & run new tasks
tsk = p$task_add(name        = "new_task",
                 description = "new task",
                 app         = "demo/tcga-demo/rna-seq-alignment-star/0",
                 inputs      = list(...))
tsk$run()
tsk$abort()
```

RUN TASKS IN BATCH MODE

```
# Batch by items
tsk = p$task_add(name        = "RNA DE Report Batch 1",
                 description = "RNA DE Analysis Report",
                 app         = rna_app$id,
                 batch       = batch(input = "bamfiles"),
                 inputs      = list(...))

# Batch by metadata
tsk = p$task_add(name        = "RNA DE Report Batch 2",
                 description = "RNA DE Analysis Report",
                 app         = rna_app$id,
                 batch       = batch(input = "fastq",
                                     c("metadata.sample_id",
                                       "metadata.library_id")),
                 inputs = list(...))
```

CONTAINERS + CWL MAKES IT EASY TO PUT NEW TOOLS ON THE CGC… AND OTHER PLACES.



+

COMMON WORKFLOW LANGUAGE

CANCER GENOMICS CLOUD
SEVEN BRIDGES

COMMON WORKFLOW LANGUAGE

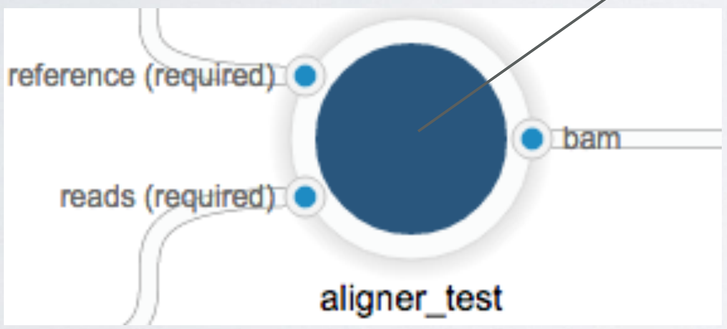
Specification for describing scalable, portable and reproducible computational workflows.

CANCER GENOMICS CLOUD
SEVEN BRIDGES

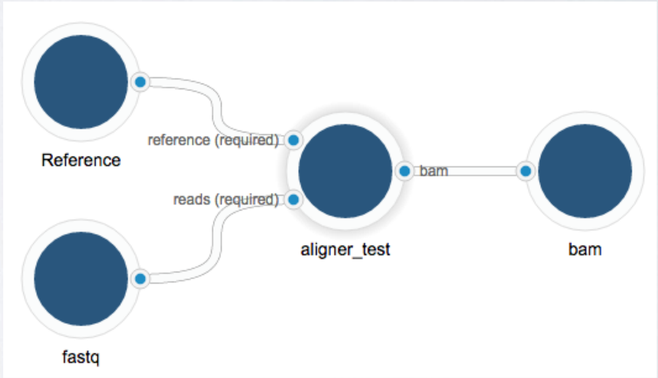
FOR DEVELOPERS, THIS MEANS…

- Easier and faster to deploy your tools.
- Write once, runs everywhere, regardless of operating system/infrastructure.

CWL TOOL: INPUTS, OUTPUTS, AND PARAMETERS

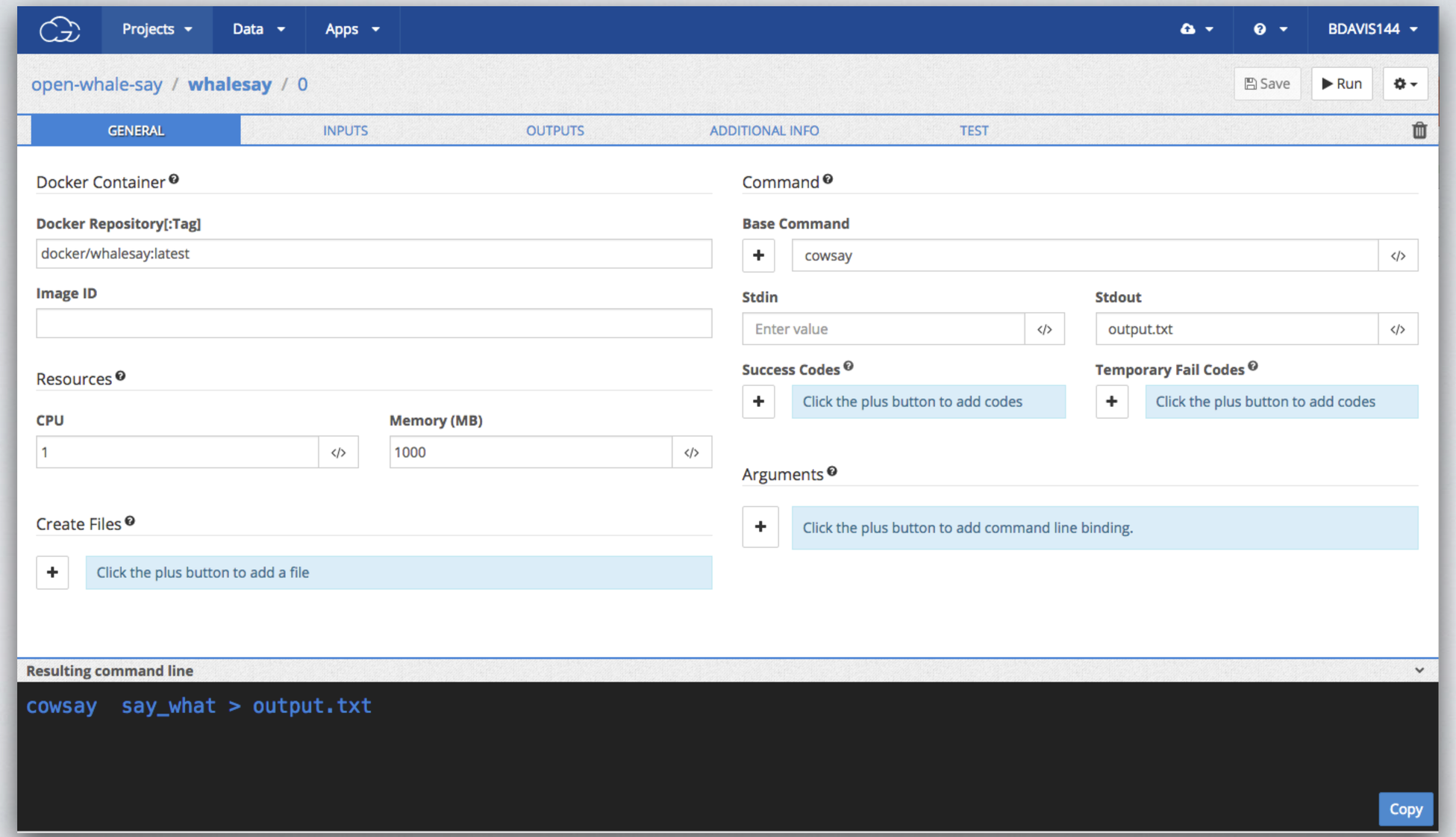



```
"softwareDescription": {
  "name": "myaligner",
  "description": "Aligns reads to a reference"
},
"documentAuthor": "kaushikghose@sbgenomics.com",
"requirements": {
  "environment": {
    "container": {
      "type": "docker",
      "uri": "",
      "imageId": ""
    }
  },
  "resources": {
    "cpu": 0,
    "mem": 5000,
    "ports": [],
    "diskSpace": 0,
    "network": false
  }
},
```



CANCER GENOMICS CLOUD
SEVEN BRIDGES

WRITE CWL WITH WEB IDE



BUILD CWL TOOLS WITH R

```
# Create CWL Tool
runif = Tool(id          = "runif",
             label       = "runif",
             hints       = requirements(docker(pull = "rocker/r-base")),
             baseCommand = "Rscript runif.R",
             stdout      = "output.txt",
             outputs     = output(id = "random", glob = "*.txt"))

# Convert to JSON or YAML
runif$toJSON()
runif$toYAML()
```

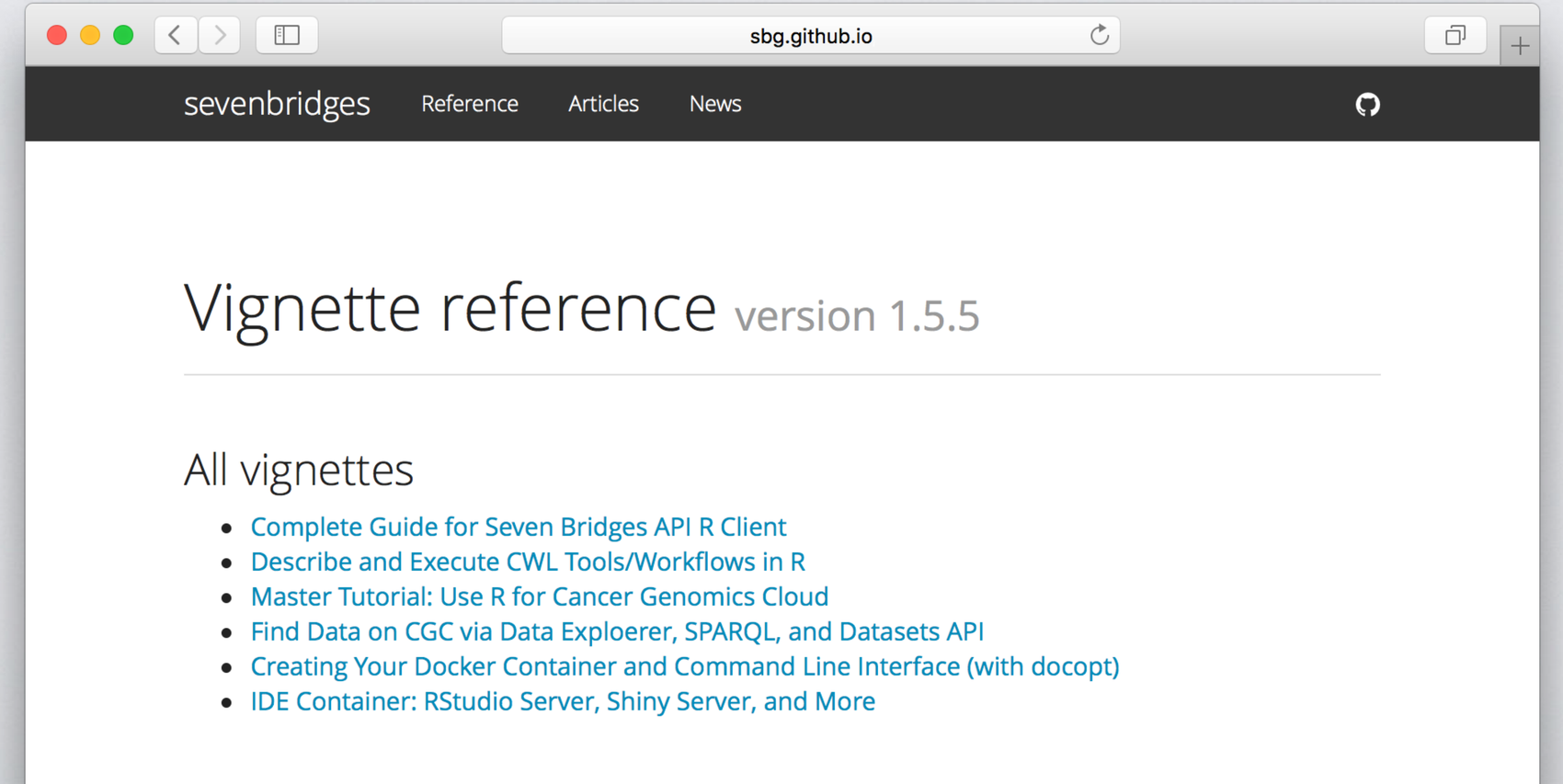
GETTING STARTED

- Try `sevenbridges` package from Bioconductor
- `docker pull sevenbridges/sevenbridges-r`
- Pull requests: github.com/sbg/sevenbridges-r

CANCER GENOMICS CLOUD
SEVEN BRIDGES

DOCUMENTATION

https://sbg.github.io/sevenbridges-r/



NEED MORE RESOURCES?

- Apply to NIHCommons Credits Pilot for new opportunities in cloud computing & storage before Jan 16:
- https://datascience.nih.gov/commons
- https://www.commons-credit-portal.org/

CANCER GENOMICS CLOUD
SEVEN BRIDGES

MORE ON CANCER GENOMICS CLOUD AND SEVEN BRIDGES PLATFORMS

- www.cancergenomicscloud.org
- www.sevenbridges.com/tcga
- www.sevenbridges.com

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Thank you!

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.

NATIONAL CANCER INSTITUTE