ISB Cancer Genomics Cloud

NCI CBIIT Speaker Series

December 9th 2015

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Google

SRA
Enduring Values. Inspired Performance.™

ISB-CGC Team Members

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ilya Shmulevich
Sheila Reynolds
Michael Miller
Phyliss Lee
Kelly Iverson
Zack Rodebaugh
Kalle Leinonen
Abigail Hahn
Eric Downes
Roger Kramer

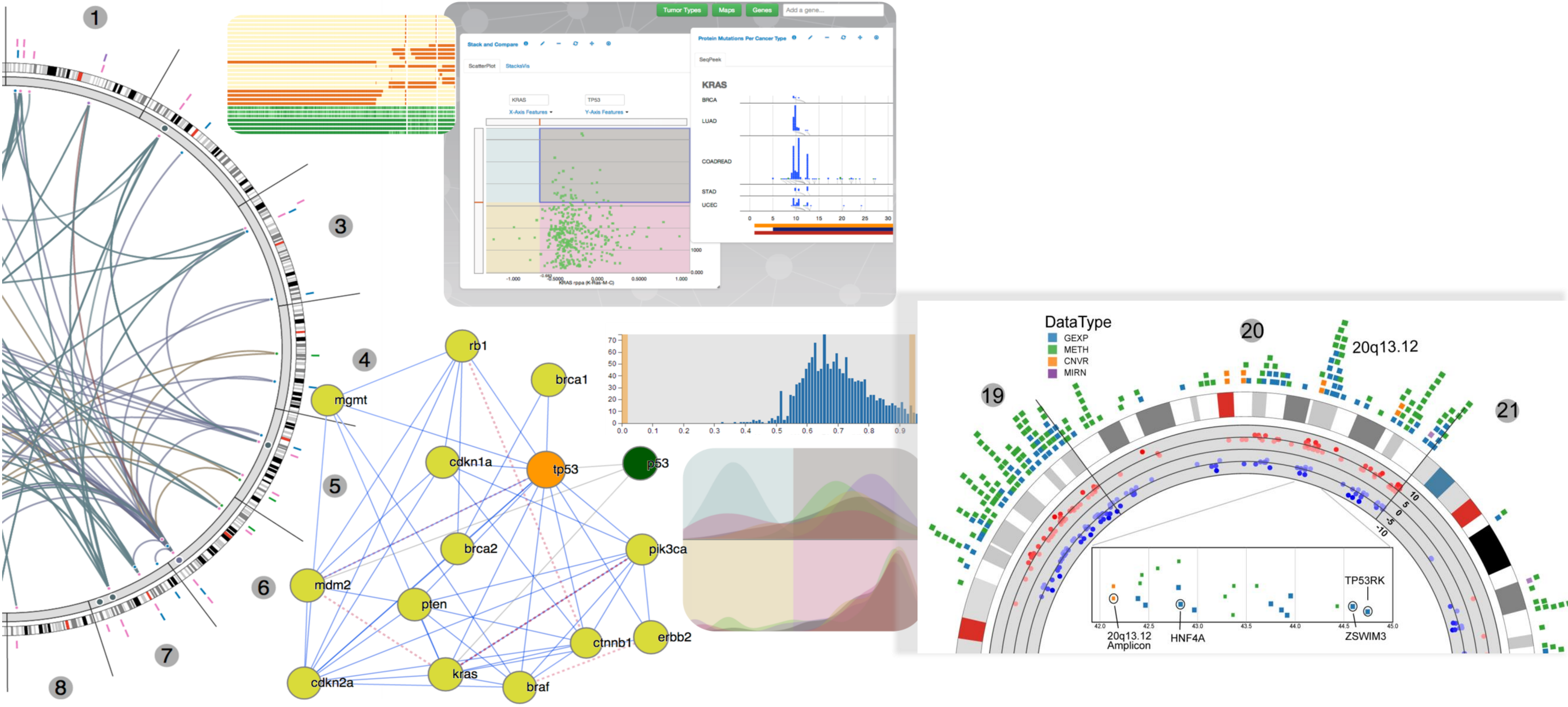
Google

Jonathan Bingham
Nicole Deflaux
Matt Bookman
Jaclyn Koller

SRA
Enduring Values. Inspired Performance.™

David Pot
Ross Casanova
Sandeep Namburi
Yan Zhang
Brian Conn

ISB GDAC in TCGA

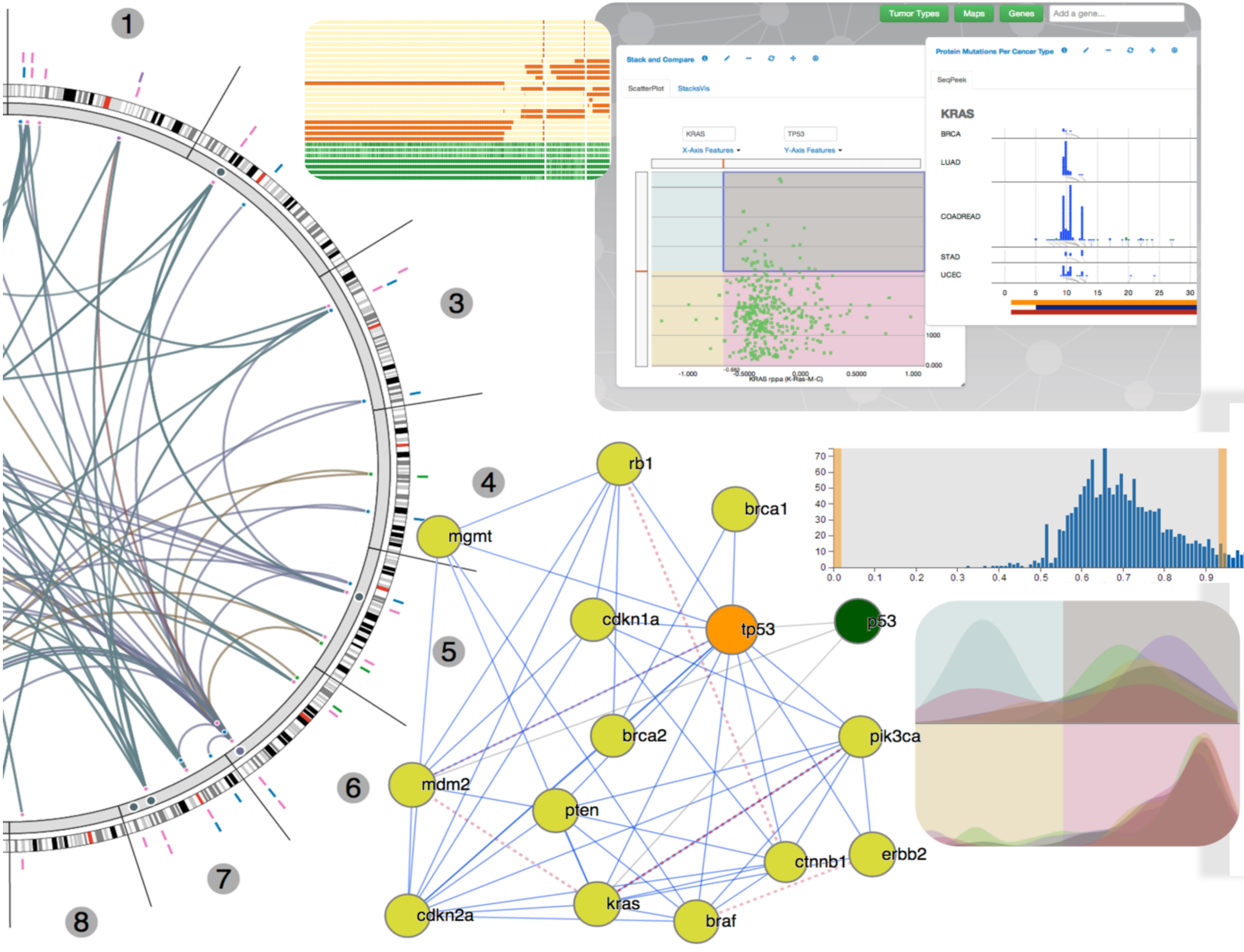


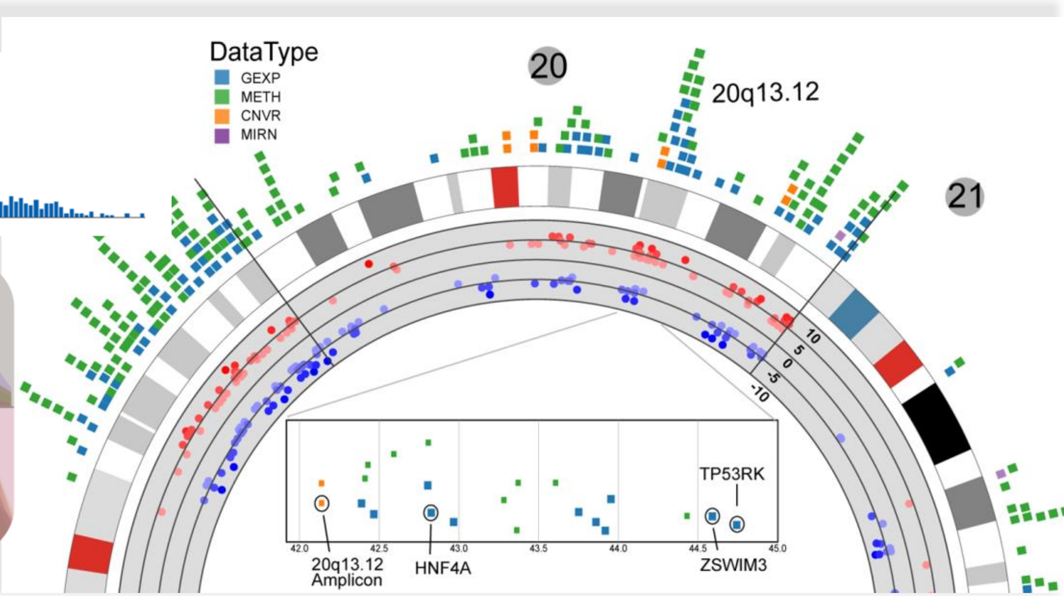
The Cancer Genome Atlas

http://explorer.cancerregulome.org

ISB GDAC in TCGA

> “[The Cloud Pilots] aim to bring data and analysis together on a single platform by creating a set of data repositories with co-located computational capacity and an Application Programming Interface (API) that provides secure data access.“

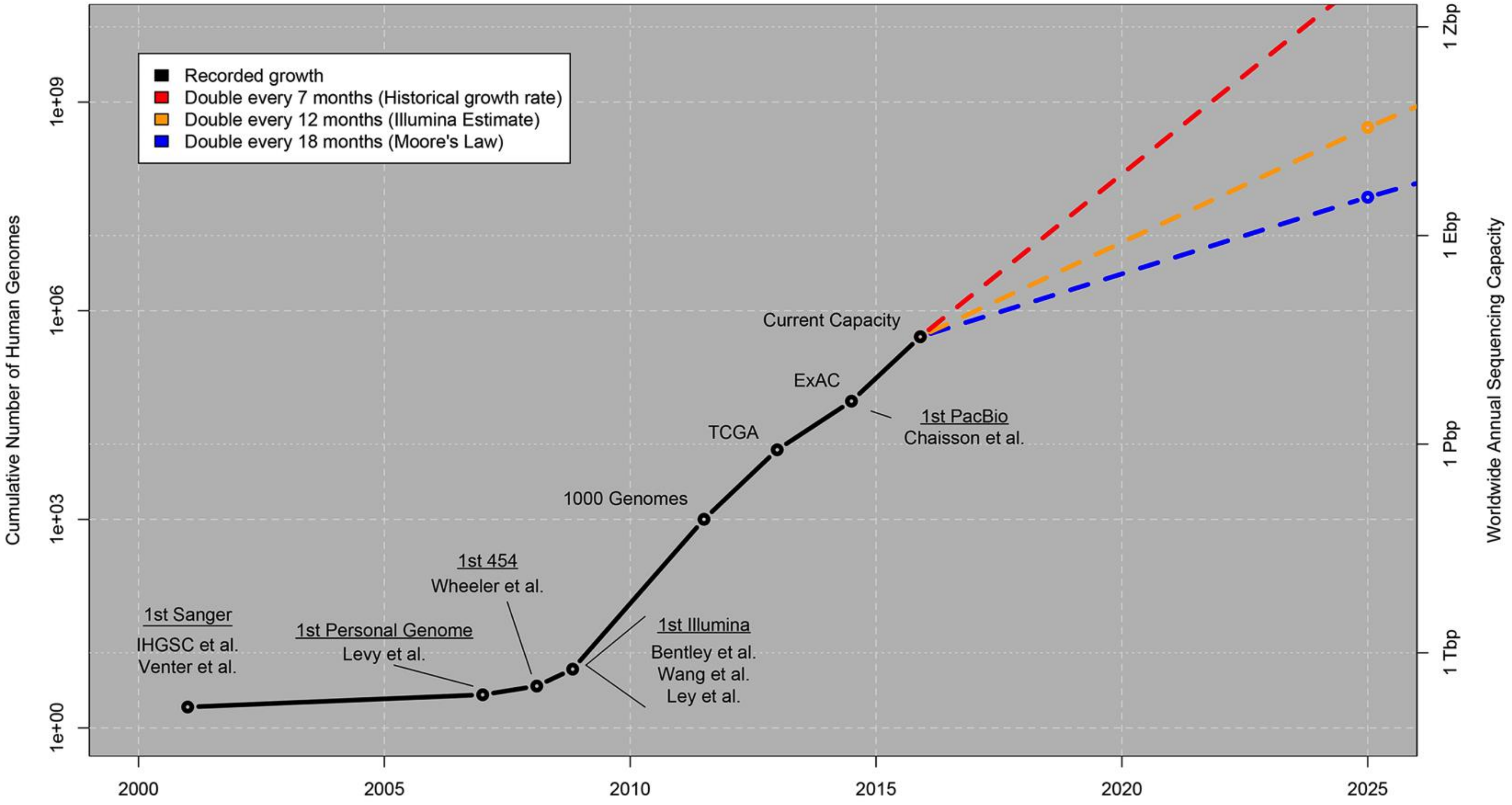




The Cancer Genome Atlas

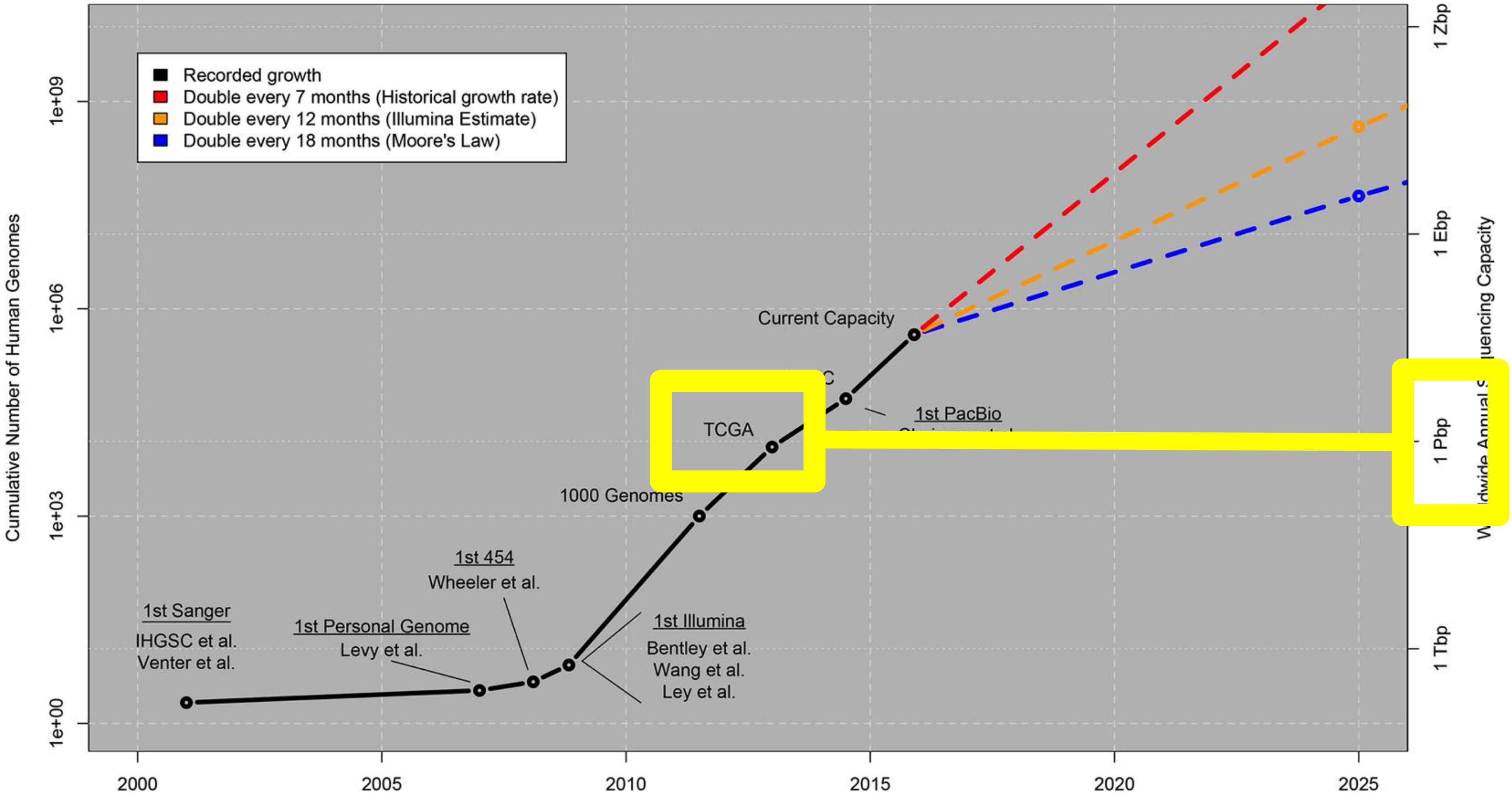
http://explorer.cancerregulome.org

The Challenge of Big Data



Big Data: Astronomical or Genomical? Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz , Saurabh Sinha , Gene E. Robinson

The Challenge of Big Data



Big Data: Astronomical or Genomical? Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz , Saurabh Sinha , Gene E. Robinson

Cloud Paradigm Shift(s)

- **Shift #1:** Move data and existing pipelines to the cloud
  - all researchers access a single copy of the data
  - everyone saves time, money, and bandwidth
  - compute-power is “near” the data
  - pay only for minutes used

- **Shift #2:** Cloud-aware computing
  - rethink/redevelop approaches to fully leverage the power of the cloud
  - massively parallel, bursty, opportunistic computing

Cloud Paradigm Shift(s)

- **Shift #1:** Move data and existing pipelines to the cloud
  - all researchers access a single copy of the data
  - everyone saves time, money, and bandwidth
  - compute-power is “near” the data
  - pay only for minutes used

- **Shift #2:** Cloud-aware computing
  - rethink/redevelop approaches to fully leverage the power of the cloud
  - massively parallel, bursty, opportunistic computing
    - *eg:* use BigQuery to calculate expression association with mutation status for **one** gene takes 7s, doing it for **all 20k** genes takes less than 9s!

The ISB Cancer Genomics Cloud

- Goals
- Approach

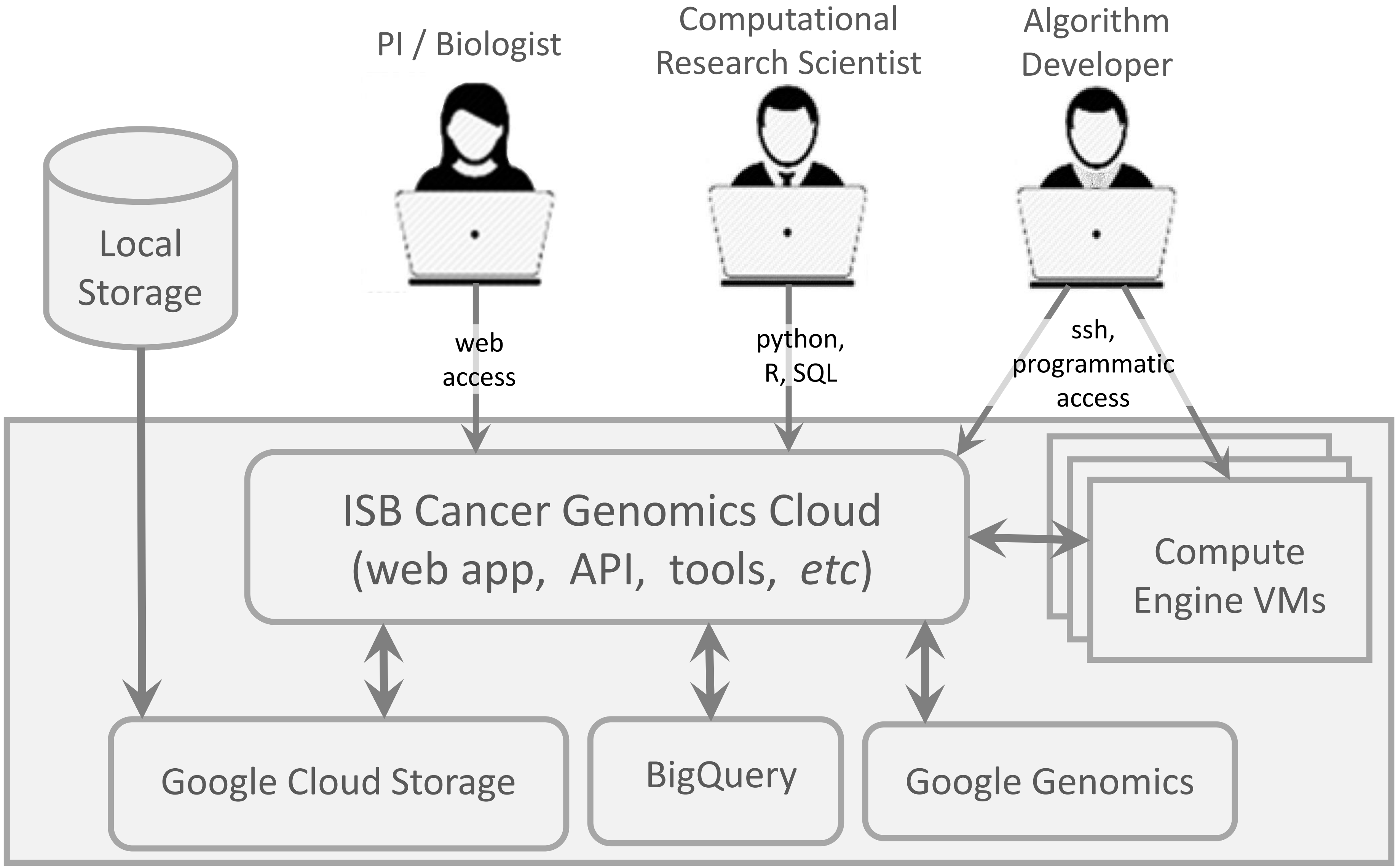
Primary Goals of the ISB-CGC

to make TCGA data, together with tools and compute-power available and accessible to a broad range of users

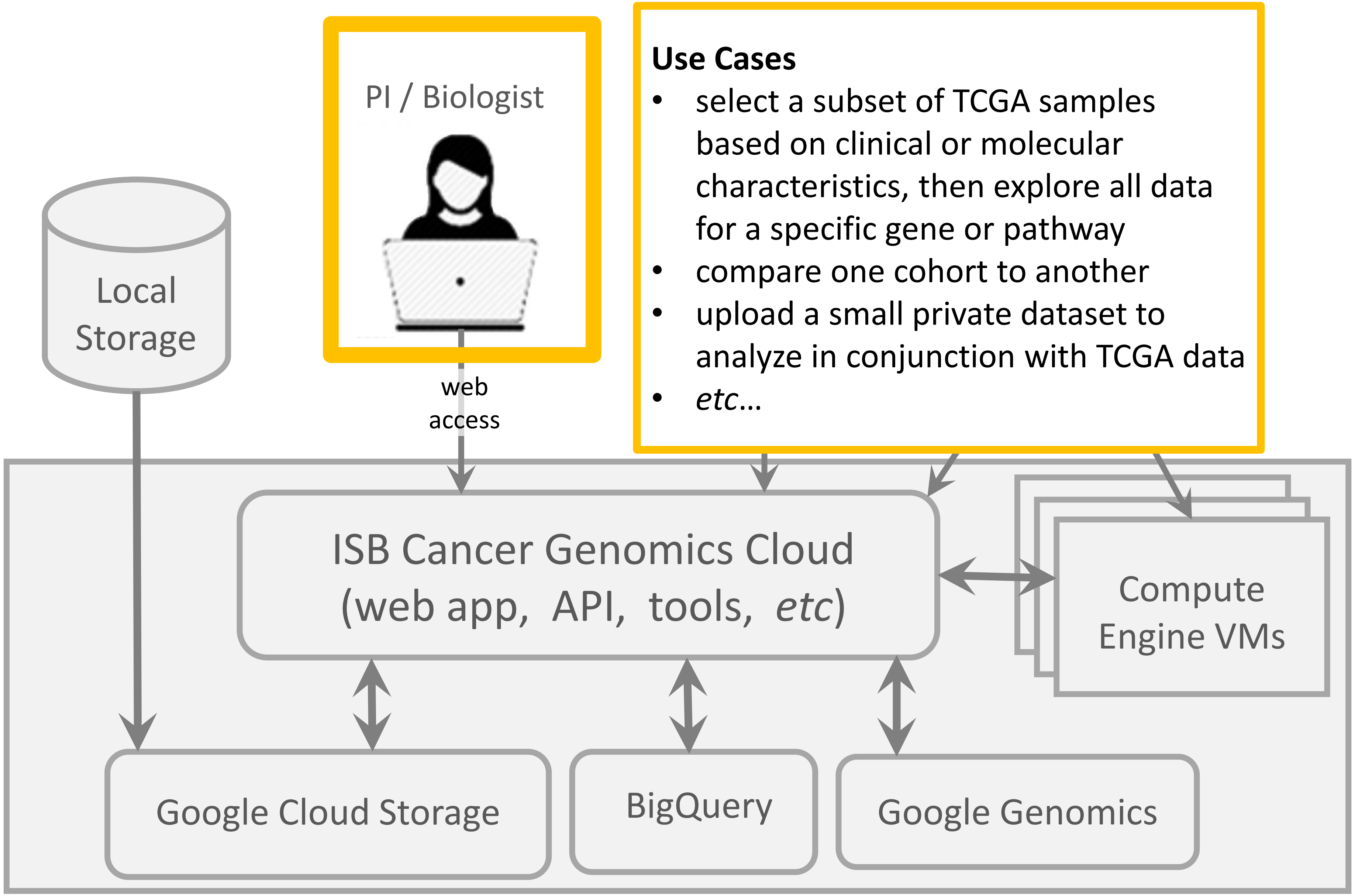
using multiple access modes:

- interactive web application
- scripting languages: R, Python, SQL
- direct programmatic access

Platform & Tools targeted to a range of users:



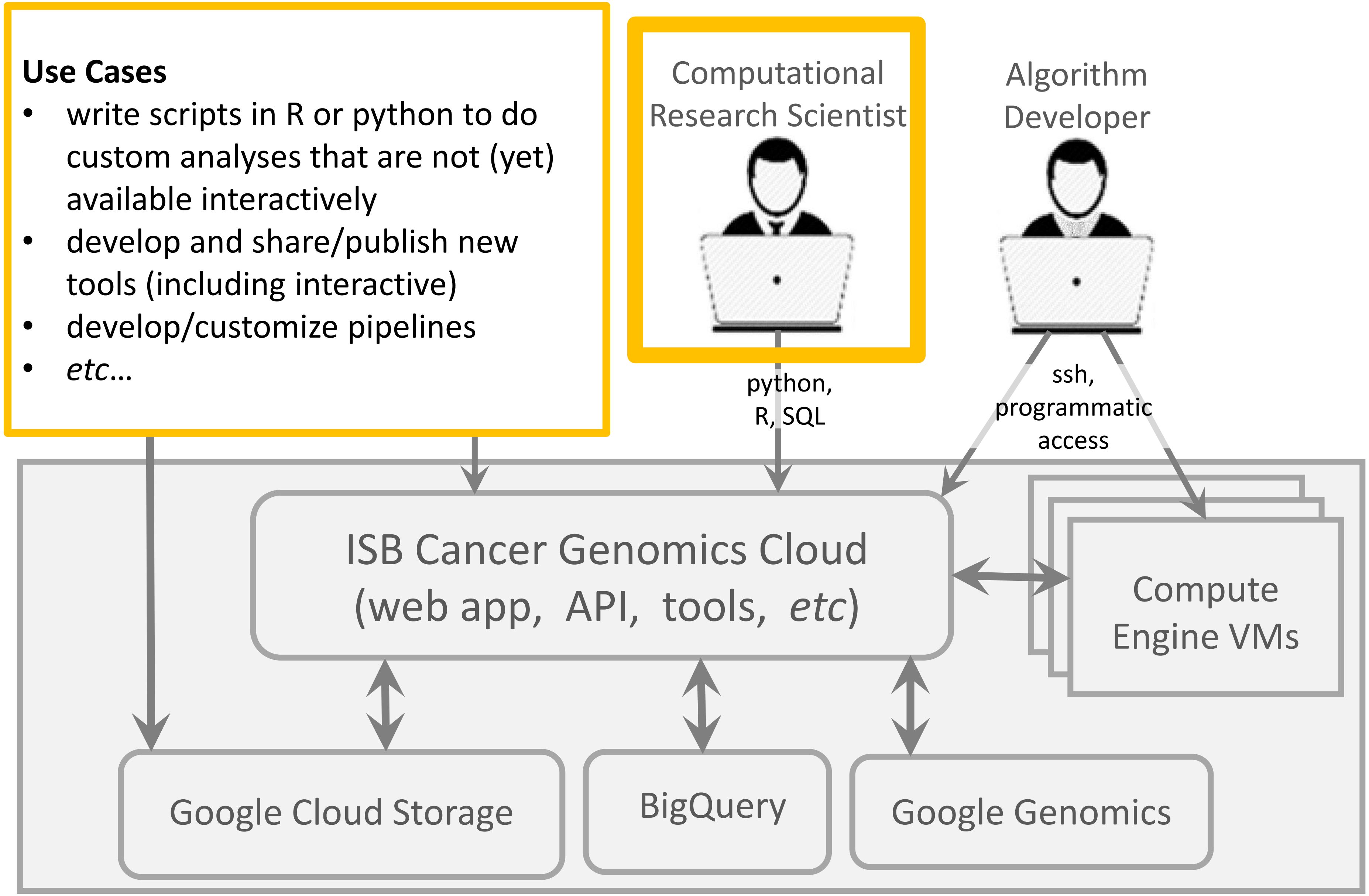
web access for the PI / Biologist:



Use Cases

- select a subset of TCGA samples based on clinical or molecular characteristics, then explore all data for a specific gene or pathway
- compare one cohort to another
- upload a small private dataset to analyze in conjunction with TCGA data
- *etc…*

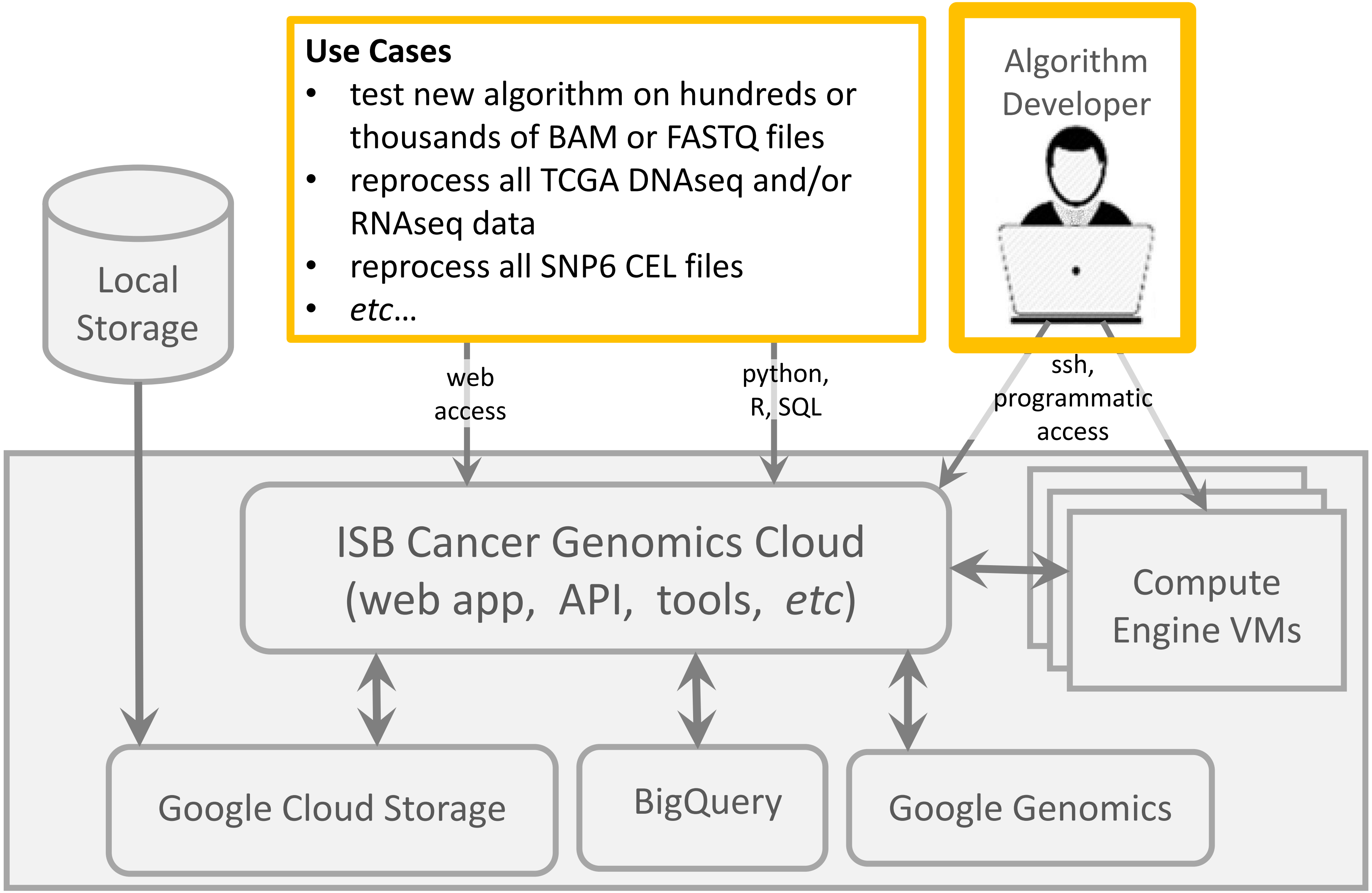
Python, R, and SQL for the Computational Scientist



Use Cases

- write scripts in R or python to do custom analyses that are not (yet) available interactively
- develop and share/publish new tools (including interactive)
- develop/customize pipelines
- *etc…*

programmatic access for the Algorithm Developer:



Use Cases

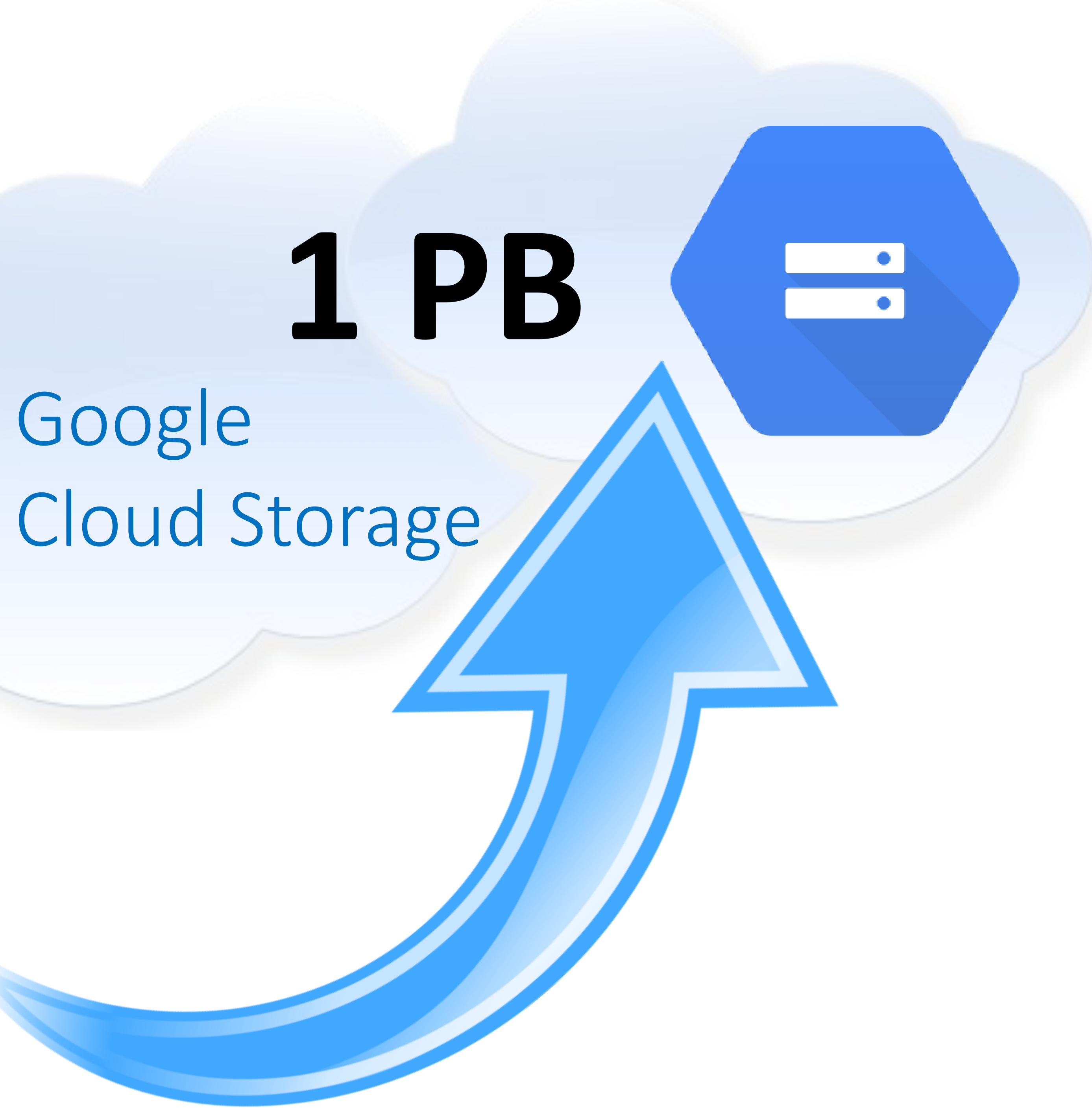
- test new algorithm on hundreds or thousands of BAM or FASTQ files
- reprocess all TCGA DNAseq and/or RNAseq data
- reprocess all SNP6 CEL files
- *etc…*

Primary Goals of the ISB-CGC

Goal #1: Data

Goal #2: Compute

Goal #1: Data



1 PB

Google Cloud Storage

THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Cancer Genomics Hub
A resource of the National Cancer Institute

1 PB

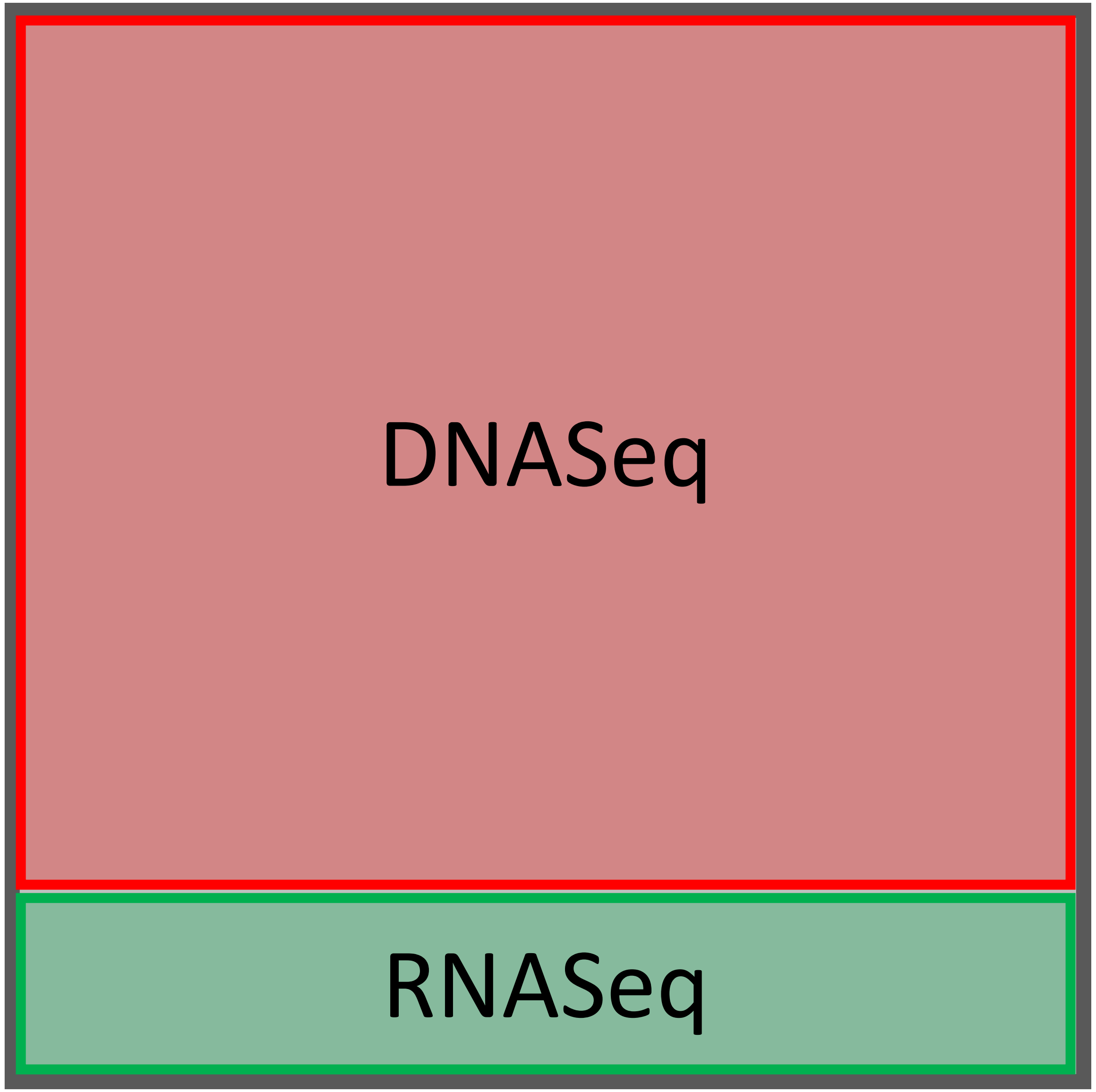
Total ***size*** of TCGA data hosted by ISB-CGC: **1 PB**

What is in there?

Total ***size*** of TCGA data hosted by ISB-CGC: **1 PB**

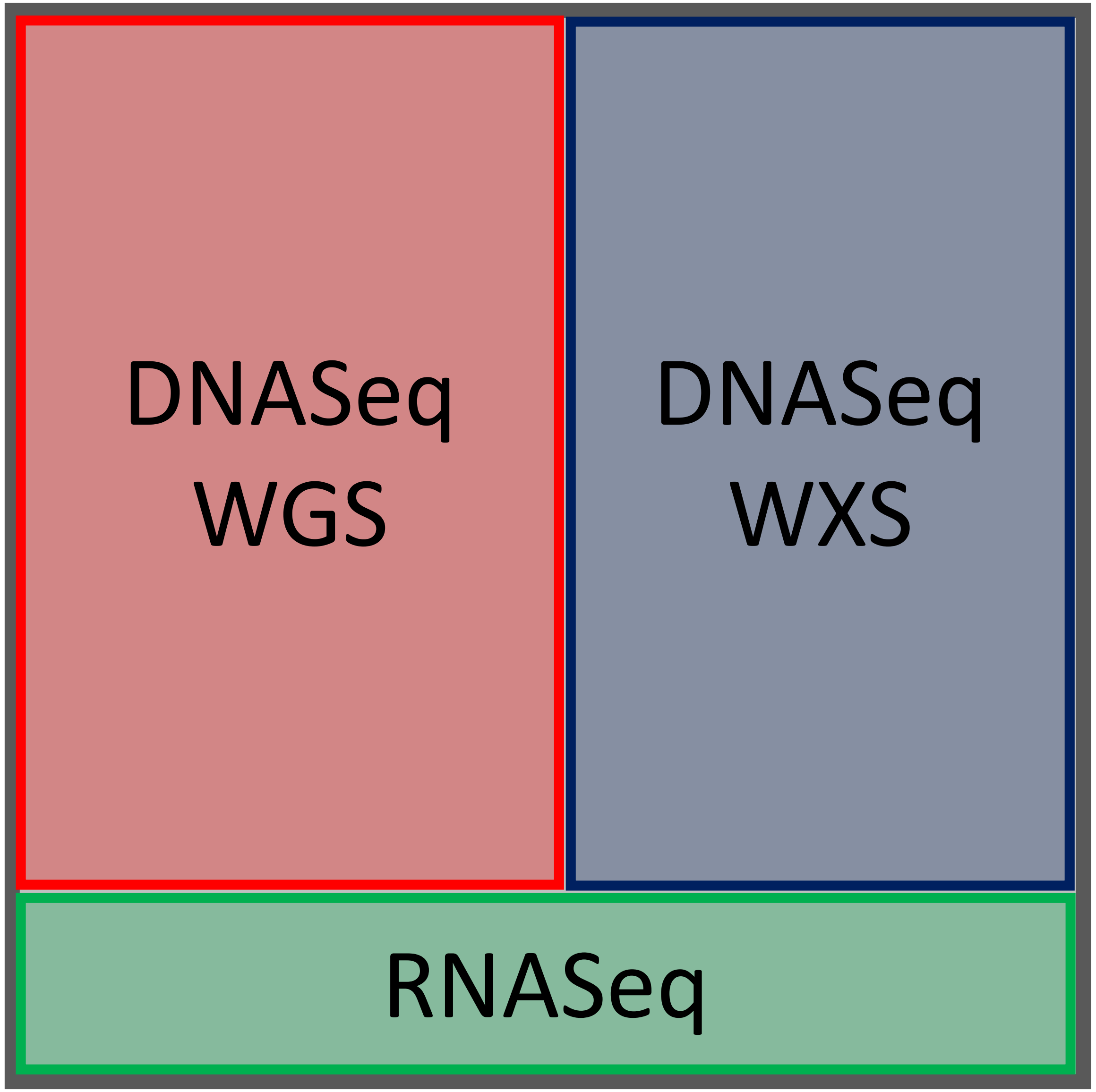
- 99.8% is low-level sequence data (Level-1)

Low-level Sequence Data



Total ***size*** of TCGA data hosted by ISB-CGC: **1 PB**

- 99.8% is low-level sequence data (Level-1)
  - 85% is DNASeq data
  - 15% is RNASeq data (including miRNAseq)

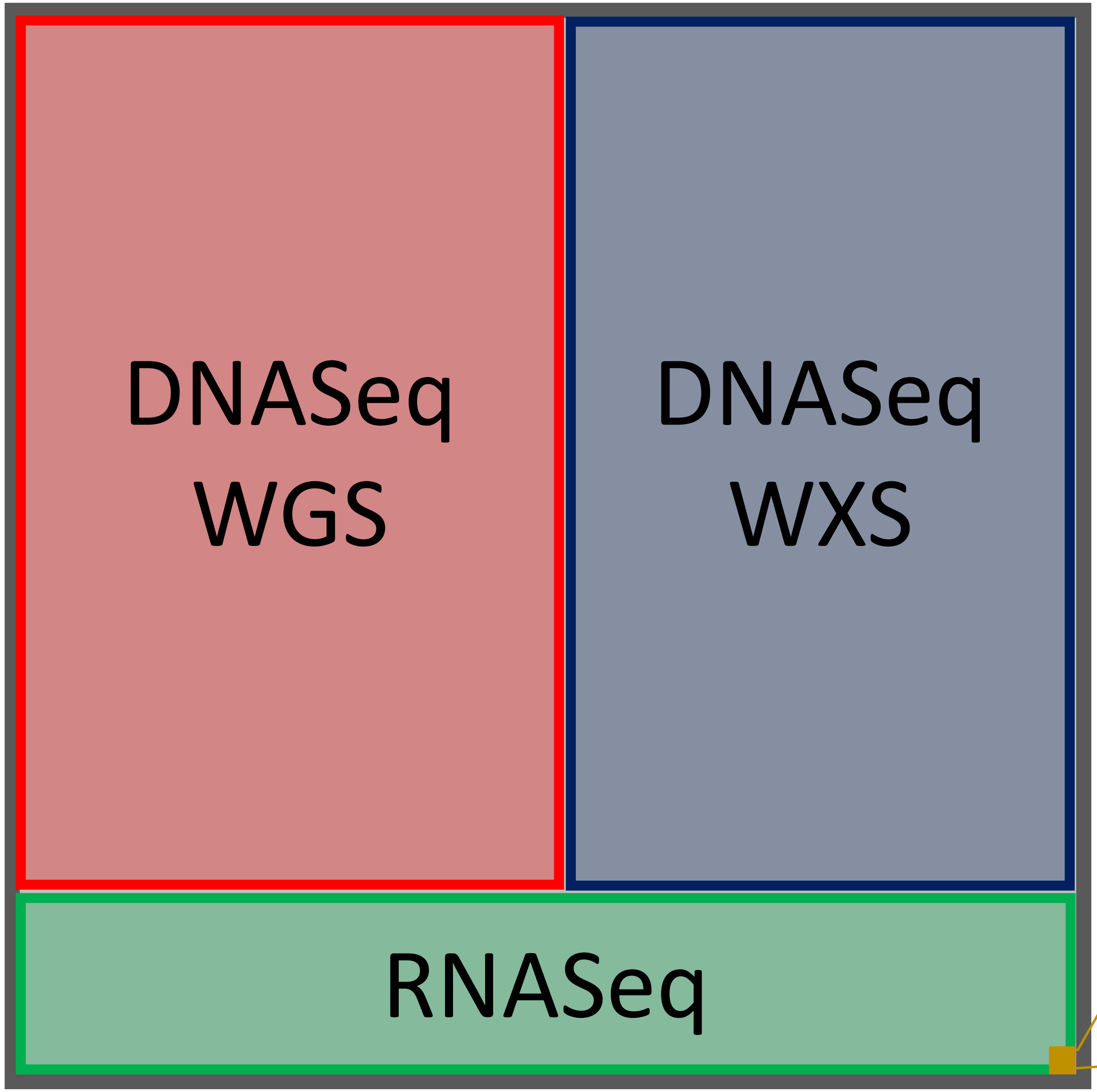


Total size of TCGA data hosted by ISB-CGC: 1 PB

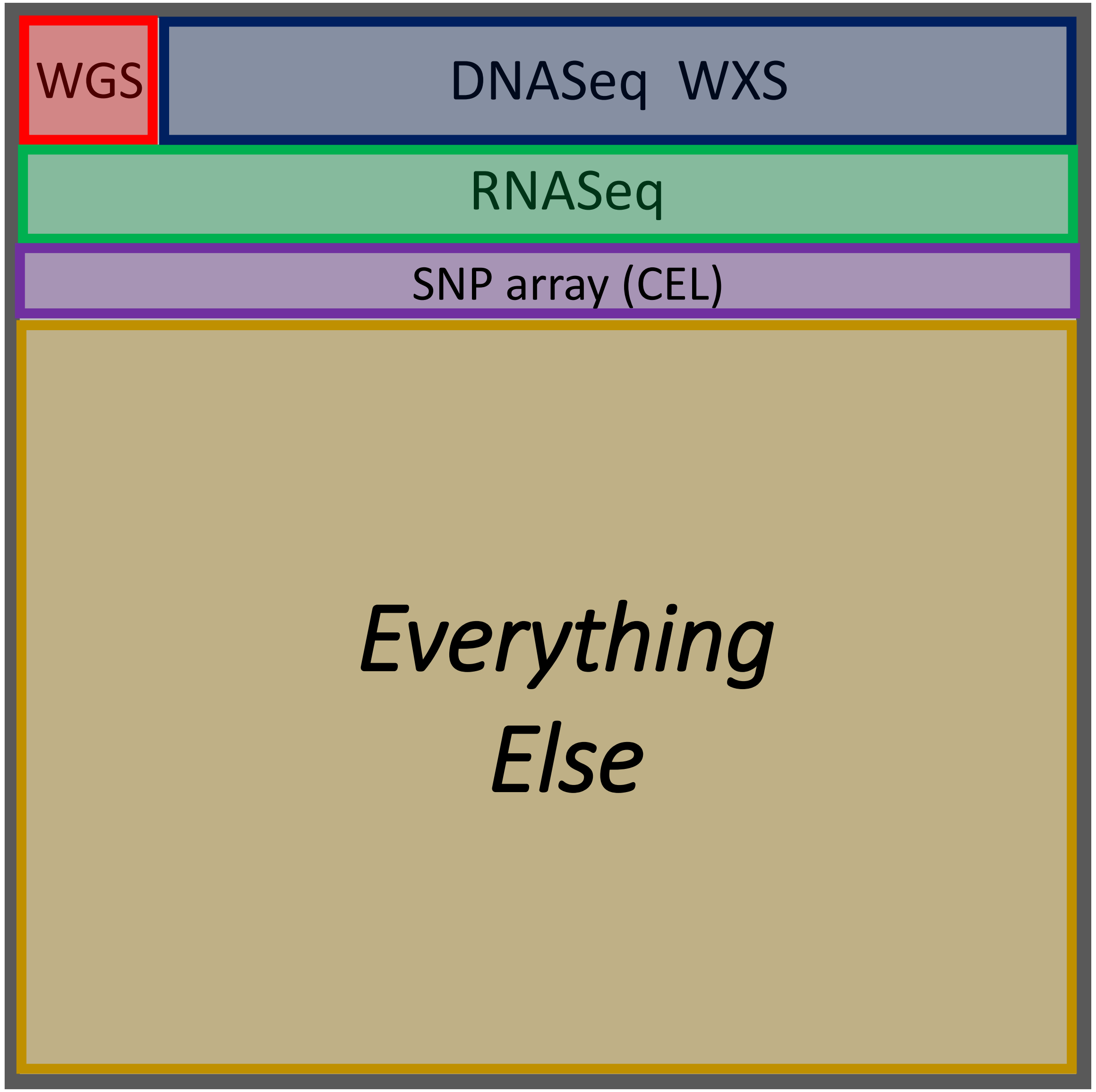
- 99.8% is low-level sequence data (Level-1)
    - 85% is DNASeq data
        - 52% is whole genome sequence
        - 48% is exome sequence
    - 15% is RNASeq data (including miRNAseq)

Total *size* of TCGA data hosted by ISB-CGC: **1 PB**

- 99.8% is low-level sequence data (Level-1)
  - 85% is DNASeq data
    - 52% is whole genome sequence
    - 48% is exome sequence
  - 15% is RNASeq data (including miRNAseq)

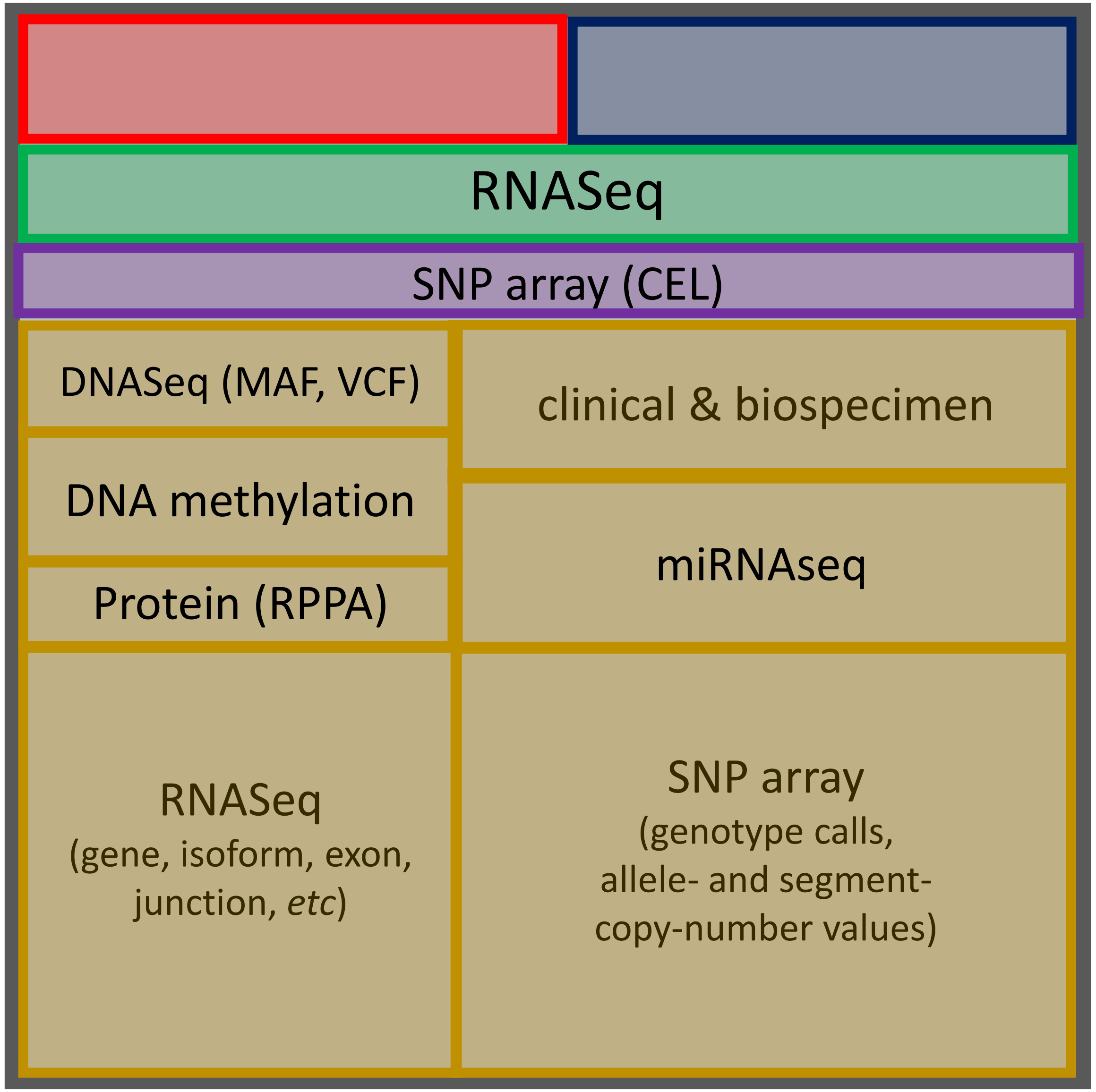


- 0.15% is low-level SNP array data (CEL files)
- 0.05% is ***all*** other data (Level-3, clinical, *etc*)



Total ***number*** of TCGA files hosted by ISB-CGC: **340K**

- 22% is low-level sequence data (Level-1)
  - 53% is DNASeq data
    - 10% is whole genome sequence
    - 90% is exome sequence
  - 47% is RNASeq data (including miRNAseq)
- 7% is low-level SNP array data (CEL files)
- 71% is ***all*** other data (Level-3, clinical, *etc*)



Total ***number*** of TCGA files hosted by ISB-CGC: **340K**

- 22% is low-level sequence data (Level-1)
  - 53% is DNASeq data
    - 10% is whole genome sequence
    - 90% is exome sequence
  - 47% is RNASeq data (including miRNAseq)
- 7% is low-level SNP array data (CEL files)
- 71% is ***all*** other data (Level-3, clinical, *etc*)

Goal #1: Data

ISB-CGC Phase 1

- Low-level sequence and SNP array data as ***files*** in **Cloud Storage**
- High-level data and annotations as ***tables*** in **BigQuery**

ISB-CGC Phase 2

- Low-level sequence data in **Google Genomics** (backed by Bigtable)
- Variant calls in **Google Genomics** and **BigQuery**

Goal #1: Data

ISB-CGC Phase 1

- Low-level sequence and SNP array data as ***files*** in **Cloud Storage**
- High-level data and annotations as ***tables*** in **BigQuery**

ISB-CGC Phase 2

- Low-level sequence data in **Google Genomics**
- Variant calls in **Google Genomics** and **BigQuery**

- BigQuery: massively parallel analytics engine pushes queries out to thousands of machines and aggregates results in seconds
- Google Genomics: read- and variant-optimized platform, supports the industry standard GA4GH API and can handle petabytes of data

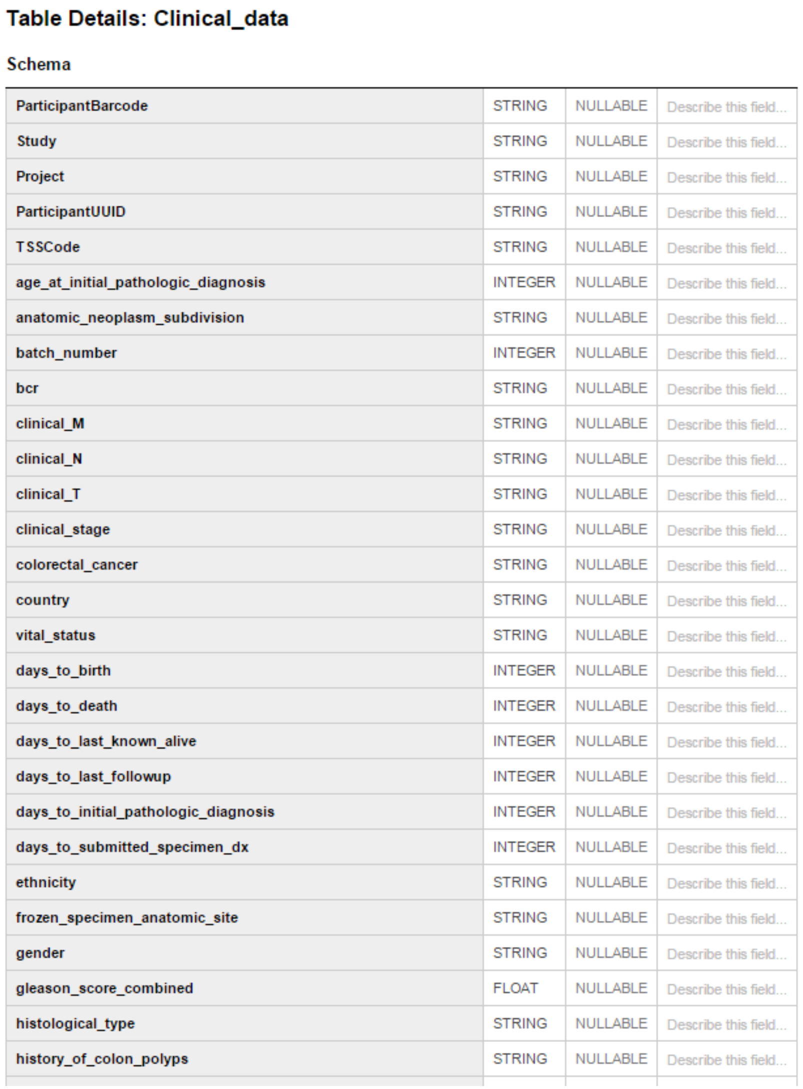


Table Details: Clinical_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
ParticipantUUID	STRING	NULLABLE	Describe this field...
TSSCode	STRING	NULLABLE	Describe this field...
age_at_initial_pathologic_diagnosis	INTEGER	NULLABLE	Describe this field...
anatomic_neoplasm_subdivision	STRING	NULLABLE	Describe this field...
batch_number	INTEGER	NULLABLE	Describe this field...
bcr	STRING	NULLABLE	Describe this field...
clinical_M	STRING	NULLABLE	Describe this field...
clinical_N	STRING	NULLABLE	Describe this field...
clinical_T	STRING	NULLABLE	Describe this field...
clinical_stage	STRING	NULLABLE	Describe this field...
colorectal_cancer	STRING	NULLABLE	Describe this field...
country	STRING	NULLABLE	Describe this field...
vital_status	STRING	NULLABLE	Describe this field...
days_to_birth	INTEGER	NULLABLE	Describe this field...
days_to_death	INTEGER	NULLABLE	Describe this field...
days_to_last_known_alive	INTEGER	NULLABLE	Describe this field...
days_to_last_followup	INTEGER	NULLABLE	Describe this field...
days_to_initial_pathologic_diagnosis	INTEGER	NULLABLE	Describe this field...
days_to_submitted_specimen_dx	INTEGER	NULLABLE	Describe this field...
ethnicity	STRING	NULLABLE	Describe this field...
frozen_specimen_anatomic_site	STRING	NULLABLE	Describe this field...
gender	STRING	NULLABLE	Describe this field...
gleason_score_combined	FLOAT	NULLABLE	Describe this field...
histological_type	STRING	NULLABLE	Describe this field...
history_of_colon_polyps	STRING	NULLABLE	Describe this field...

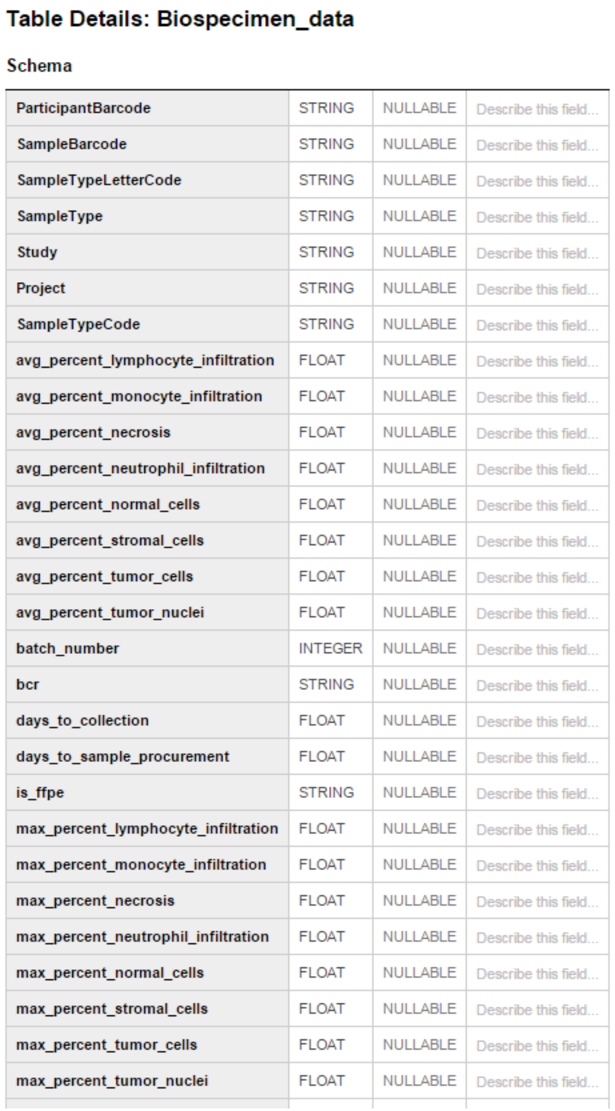


Table Details: Biospecimen_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
SampleType	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
SampleTypeCode	STRING	NULLABLE	Describe this field...
avg_percent_lymphocyte_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_monocyte_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_necrosis	FLOAT	NULLABLE	Describe this field...
avg_percent_neutrophil_infiltration	FLOAT	NULLABLE	Describe this field...
avg_percent_normal_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_stromal_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_tumor_cells	FLOAT	NULLABLE	Describe this field...
avg_percent_tumor_nuclei	FLOAT	NULLABLE	Describe this field...
batch_number	INTEGER	NULLABLE	Describe this field...
bcr	STRING	NULLABLE	Describe this field...
days_to_collection	FLOAT	NULLABLE	Describe this field...
days_to_sample_procurement	FLOAT	NULLABLE	Describe this field...
is_ffpe	STRING	NULLABLE	Describe this field...
max_percent_lymphocyte_infiltration	FLOAT	NULLABLE	Describe this field...
max_percent_monocyte_infiltration	FLOAT	NULLABLE	Describe this field...
max_percent_necrosis	FLOAT	NULLABLE	Describe this field...
max_percent_neutrophil_infiltration	FLOAT	NULLABLE	Describe this field...
max_percent_normal_cells	FLOAT	NULLABLE	Describe this field...
max_percent_stromal_cells	FLOAT	NULLABLE	Describe this field...
max_percent_tumor_cells	FLOAT	NULLABLE	Describe this field...
max_percent_tumor_nuclei	FLOAT	NULLABLE	Describe this field...

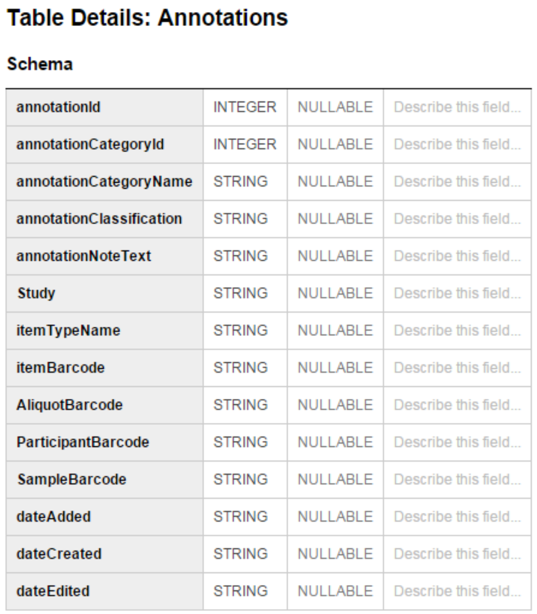


Table Details: Annotations

Schema

annotationId	INTEGER	NULLABLE	Describe this field...
annotationCategoryId	INTEGER	NULLABLE	Describe this field...
annotationCategoryName	STRING	NULLABLE	Describe this field...
annotationClassification	STRING	NULLABLE	Describe this field...
annotationNoteText	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
itemTypeName	STRING	NULLABLE	Describe this field...
itemBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
dateAdded	STRING	NULLABLE	Describe this field...
dateCreated	STRING	NULLABLE	Describe this field...
dateEdited	STRING	NULLABLE	Describe this field...

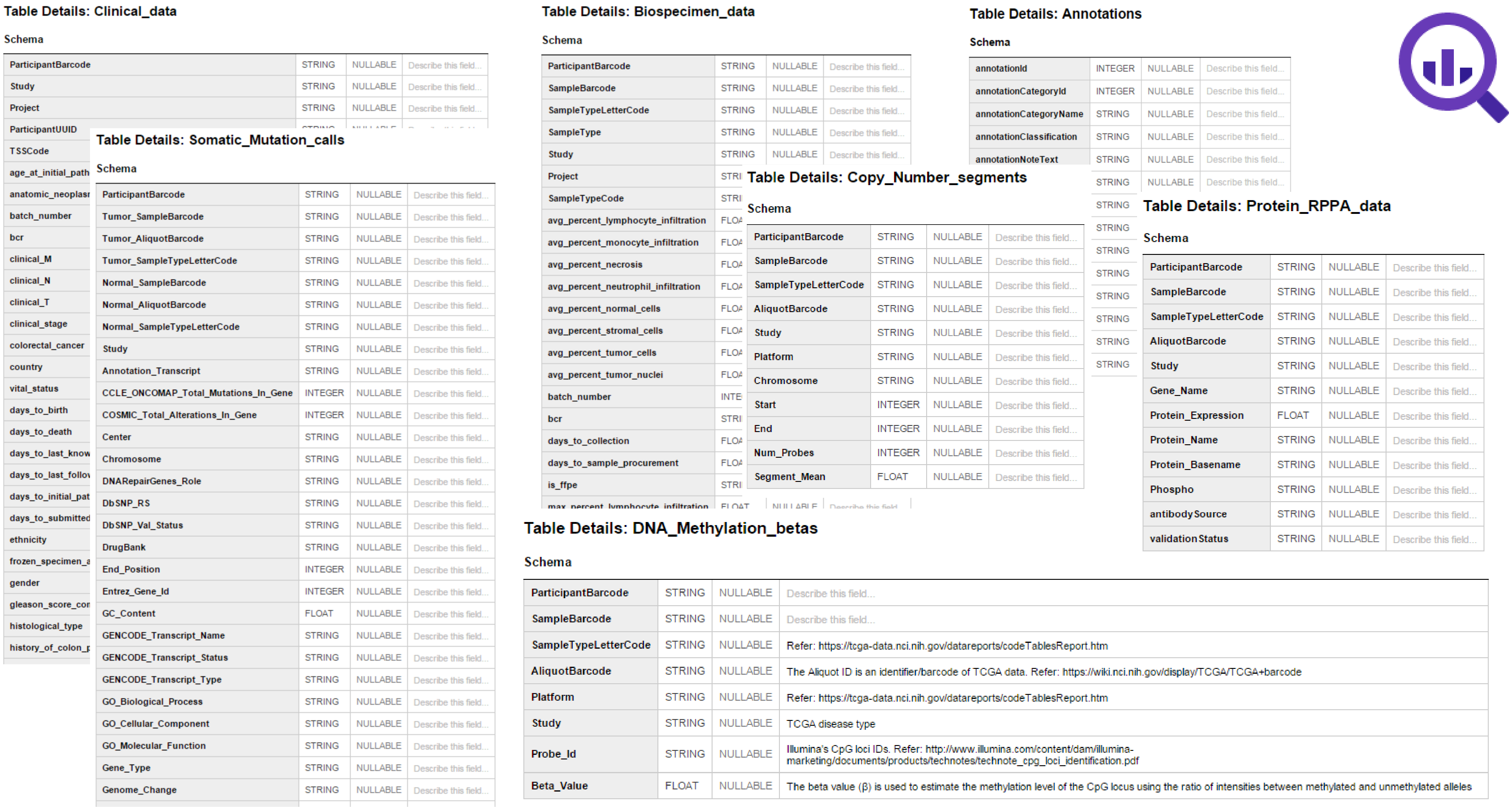


Table Details: Clinical_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
ParticipantUUID	STRING	NULLABLE	Describe this field...
TSSCode			
age_at_initial_path			
anatomic_neoplasr			
batch_number			
bcr			
clinical_M			
clinical_N			
clinical_T			
clinical_stage			
colorectal_cancer			
country			
vital_status			
days_to_birth			
days_to_death			
days_to_last_know			
days_to_last_follov			
days_to_initial_pat			
days_to_submitted			
ethnicity			
frozen_specimen_a			
gender			
gleason_score_cor			
histological_type			
history_of_colon_p			

Table Details: Somatic_Mutation_calls

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleBarcode	STRING	NULLABLE	Describe this field...
Tumor_AliquotBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Normal_SampleBarcode	STRING	NULLABLE	Describe this field...
Normal_AliquotBarcode	STRING	NULLABLE	Describe this field...
Normal_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Annotation_Transcript	STRING	NULLABLE	Describe this field...
CCLE_ONCOMAP_Total_Mutations_In_Gene	INTEGER	NULLABLE	Describe this field...
COSMIC_Total_Alterations_In_Gene	INTEGER	NULLABLE	Describe this field...
Center	STRING	NULLABLE	Describe this field...
Chromosome	STRING	NULLABLE	Describe this field...
DNARepairGenes_Role	STRING	NULLABLE	Describe this field...
DbSNP_RS	STRING	NULLABLE	Describe this field...
DbSNP_Val_Status	STRING	NULLABLE	Describe this field...
DrugBank	STRING	NULLABLE	Describe this field...
End_Position	INTEGER	NULLABLE	Describe this field...
Entrez_Gene_Id	INTEGER	NULLABLE	Describe this field...
GC_Content	FLOAT	NULLABLE	Describe this field...
GENCODE_Transcript_Name	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Status	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Type	STRING	NULLABLE	Describe this field...
GO_Biological_Process	STRING	NULLABLE	Describe this field...
GO_Cellular_Component	STRING	NULLABLE	Describe this field...
GO_Molecular_Function	STRING	NULLABLE	Describe this field...
Gene_Type	STRING	NULLABLE	Describe this field...
Genome_Change	STRING	NULLABLE	Describe this field...

Table Details: Biospecimen_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
SampleType	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRI		
SampleTypeCode	STRI		
avg_percent_lymphocyte_infiltration	FLOA		
avg_percent_monocyte_infiltration	FLOA		
avg_percent_necrosis	FLOA		
avg_percent_neutrophil_infiltration	FLOA		
avg_percent_normal_cells	FLOA		
avg_percent_stromal_cells	FLOA		
avg_percent_tumor_cells	FLOA		
avg_percent_tumor_nuclei	FLOA		
batch_number	INTE		
bcr	STRI		
days_to_collection	FLOA		
days_to_sample_procurement	FLOA		
is_ffpe	STRI		
max_percent_lymphocyte_infiltration	FLOAT	NULLABLE	Describe this field...

Table Details: Annotations

Schema

annotationId	INTEGER	NULLABLE	Describe this field...
annotationCategoryId	INTEGER	NULLABLE	Describe this field...
annotationCategoryName	STRING	NULLABLE	Describe this field...
annotationClassification	STRING	NULLABLE	Describe this field...
annotationNoteText	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		

Table Details: Copy_Number_segments

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
Chromosome	STRING	NULLABLE	Describe this field...
Start	INTEGER	NULLABLE	Describe this field...
End	INTEGER	NULLABLE	Describe this field...
Num_Probes	INTEGER	NULLABLE	Describe this field...
Segment_Mean	FLOAT	NULLABLE	Describe this field...

Table Details: Protein_RPPA_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Gene_Name	STRING	NULLABLE	Describe this field...
Protein_Expression	FLOAT	NULLABLE	Describe this field...
Protein_Name	STRING	NULLABLE	Describe this field...
Protein_Basename	STRING	NULLABLE	Describe this field...
Phospho	STRING	NULLABLE	Describe this field...
antibodySource	STRING	NULLABLE	Describe this field...
validationStatus	STRING	NULLABLE	Describe this field...

Table Details: DNA_Methylation_betas

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Refer: https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm
AliquotBarcode	STRING	NULLABLE	The Aliquot ID is an identifier/barcode of TCGA data. Refer: https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode
Platform	STRING	NULLABLE	Refer: https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm
Study	STRING	NULLABLE	TCGA disease type
Probe_Id	STRING	NULLABLE	Illumina's CpG loci IDs. Refer: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_cpg_loci_identification.pdf
Beta_Value	FLOAT	NULLABLE	The beta value (β) is used to estimate the methylation level of the CpG locus using the ratio of intensities between methylated and unmethylated alleles

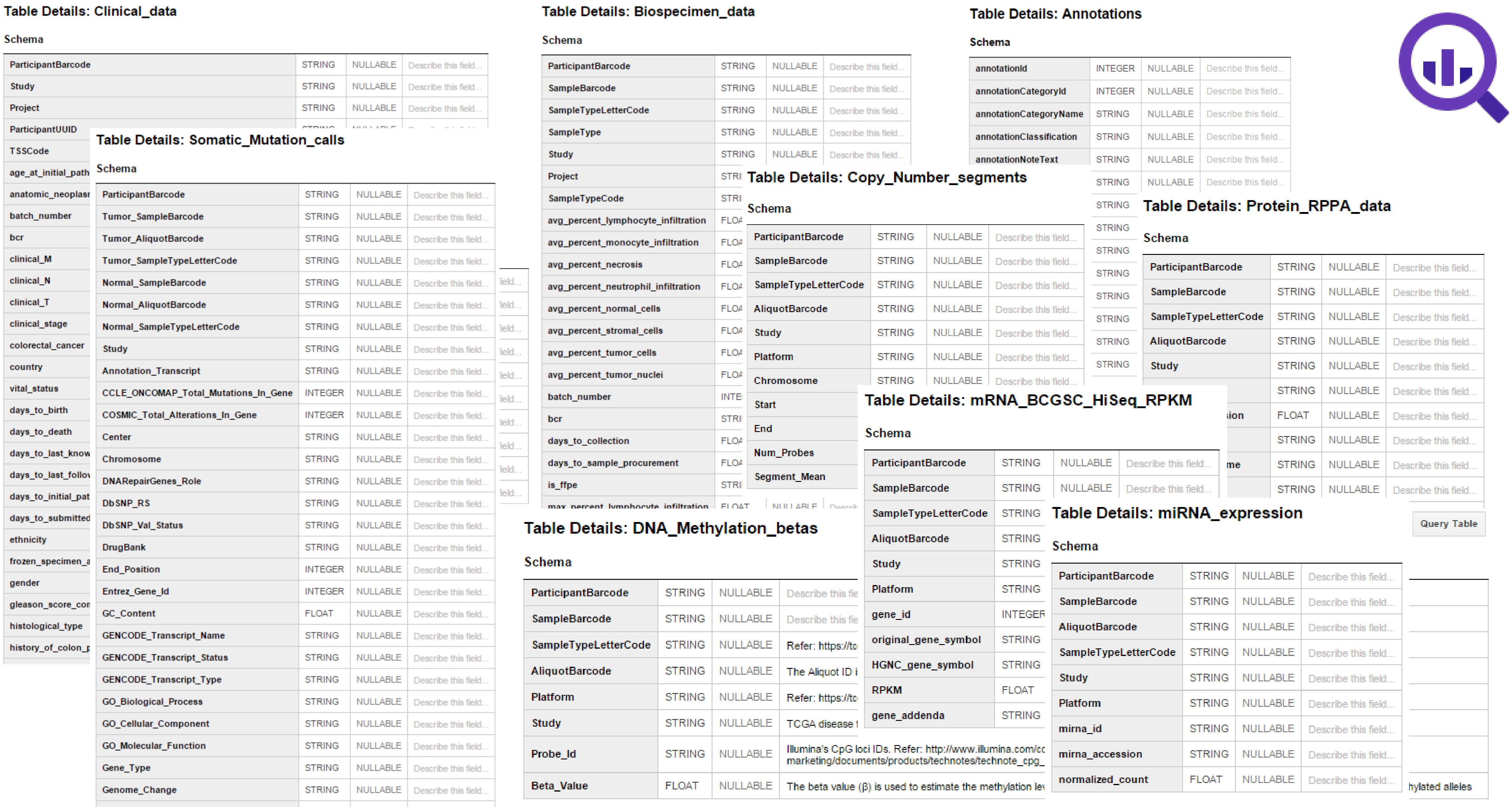


Table Details: Clinical_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRING	NULLABLE	Describe this field...
ParticipantUUID			
TSSCode			
age_at_initial_path			
anatomic_neoplasr			
batch_number			
bcr			
clinical_M			
clinical_N			
clinical_T			
clinical_stage			
colorectal_cancer			
country			
vital_status			
days_to_birth			
days_to_death			
days_to_last_know			
days_to_last_follov			
days_to_initial_pat			
days_to_submitted			
ethnicity			
frozen_specimen_a			
gender			
gleason_score_cor			
histological_type			
history_of_colon_p			

Table Details: Somatic_Mutation_calls

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleBarcode	STRING	NULLABLE	Describe this field...
Tumor_AliquotBarcode	STRING	NULLABLE	Describe this field...
Tumor_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Normal_SampleBarcode	STRING	NULLABLE	Describe this field...
Normal_AliquotBarcode	STRING	NULLABLE	Describe this field...
Normal_SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Annotation_Transcript	STRING	NULLABLE	Describe this field...
CCLE_ONCOMAP_Total_Mutations_In_Gene	INTEGER	NULLABLE	Describe this field...
COSMIC_Total_Alterations_In_Gene	INTEGER	NULLABLE	Describe this field...
Center	STRING	NULLABLE	Describe this field...
Chromosome	STRING	NULLABLE	Describe this field...
DNARepairGenes_Role	STRING	NULLABLE	Describe this field...
DbSNP_RS	STRING	NULLABLE	Describe this field...
DbSNP_Val_Status	STRING	NULLABLE	Describe this field...
DrugBank	STRING	NULLABLE	Describe this field...
End_Position	INTEGER	NULLABLE	Describe this field...
Entrez_Gene_Id	INTEGER	NULLABLE	Describe this field...
GC_Content	FLOAT	NULLABLE	Describe this field...
GENCODE_Transcript_Name	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Status	STRING	NULLABLE	Describe this field...
GENCODE_Transcript_Type	STRING	NULLABLE	Describe this field...
GO_Biological_Process	STRING	NULLABLE	Describe this field...
GO_Cellular_Component	STRING	NULLABLE	Describe this field...
GO_Molecular_Function	STRING	NULLABLE	Describe this field...
Gene_Type	STRING	NULLABLE	Describe this field...
Genome_Change	STRING	NULLABLE	Describe this field...

Table Details: Biospecimen_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
SampleType	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Project	STRI		
SampleTypeCode	STRI		
avg_percent_lymphocyte_infiltration	FLOA		
avg_percent_monocyte_infiltration	FLOA		
avg_percent_necrosis	FLOA		
avg_percent_neutrophil_infiltration	FLOA		
avg_percent_normal_cells	FLOA		
avg_percent_stromal_cells	FLOA		
avg_percent_tumor_cells	FLOA		
avg_percent_tumor_nuclei	FLOA		
batch_number	INTE		
bcr	STRI		
days_to_collection	FLOA		
days_to_sample_procurement	FLOA		
is_ffpe	STRI		
max_percent_lymphocyte_infiltration	FLOAT	NULLABLE	Descri

Table Details: Annotations

Schema

annotationId	INTEGER	NULLABLE	Describe this field...
annotationCategoryId	INTEGER	NULLABLE	Describe this field...
annotationCategoryName	STRING	NULLABLE	Describe this field...
annotationClassification	STRING	NULLABLE	Describe this field...
annotationNoteText	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		
	STRING		

Table Details: Copy_Number_segments

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
Chromosome	STRING	NULLABLE	Describe this field...
Start			
End			
Num_Probes			
Segment_Mean			

Table Details: Protein_RPPA_data

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
ion	FLOAT	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
me	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...

Table Details: mRNA_BCGSC_HiSeq_RPKM

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING		
AliquotBarcode	STRING		
Study	STRING		
Platform	STRING		
gene_id	INTEGER		
original_gene_symbol	STRING		
HGNC_gene_symbol	STRING		
RPKM	FLOAT		
gene_addenda	STRING		

Table Details: DNA_Methylation_betas

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this fie
SampleBarcode	STRING	NULLABLE	Describe this fie
SampleTypeLetterCode	STRING	NULLABLE	Refer: https://tc
AliquotBarcode	STRING	NULLABLE	The Aliquot ID i
Platform	STRING	NULLABLE	Refer: https://tc
Study	STRING	NULLABLE	TCGA disease t
Probe_Id	STRING	NULLABLE	Illumina's CpG loci IDs. Refer: http://www.illumina.com/cc marketing/documents/products/technotes/technote_cpg_
Beta_Value	FLOAT	NULLABLE	The beta value (β) is used to estimate the methylation lev ... hylated alleles

Table Details: miRNA_expression

Query Table

Schema

ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
mirna_id	STRING	NULLABLE	Describe this field...
mirna_accession	STRING	NULLABLE	Describe this field...
normalized_count	FLOAT	NULLABLE	Describe this field...

Bring your data to BigQuery!

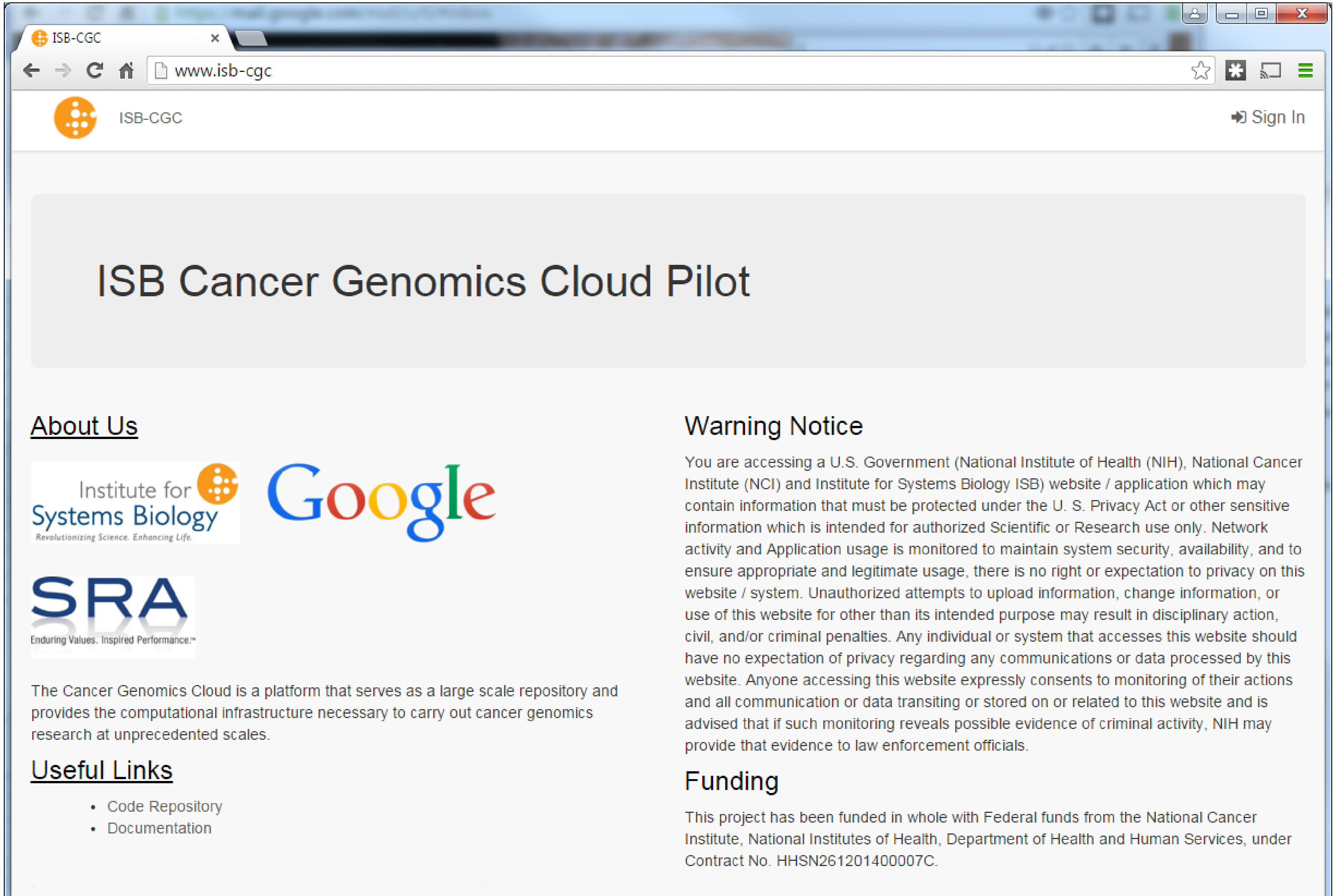
- easily integrate with other BigQuery datasets … if other people put their data and annotations into BigQuery tables
- *eg* Tute Genomics

- Let’s put out a call to researchers to make data, annotations, etc available for all to use in BigQuery!
  - TCGA Level-3 data (500 GB) -- $10 per month
  - Tute Genomics (649 GB and 8.6 billion rows) -- $13 per month
  - GENCODE (593 MB table with 2.6 million rows) -- only 14 cents per year

Goal #2: Compute

1. PI / Biologist: web-based interaction
2. Computational Research Scientist: R, Python, SQL
3. Algorithm Developer: VMs, Container Engine, Dataproc, Dataflow

web access for the PI / Biologist


ISB-CGC
www.isb-cgc
ISB-CGC
Sign In

ISB Cancer Genomics Cloud Pilot

About Us

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.
Google
SRA
Enduring Values. Inspired Performance.™

The Cancer Genomics Cloud is a platform that serves as a large scale repository and provides the computational infrastructure necessary to carry out cancer genomics research at unprecedented scales.

Useful Links

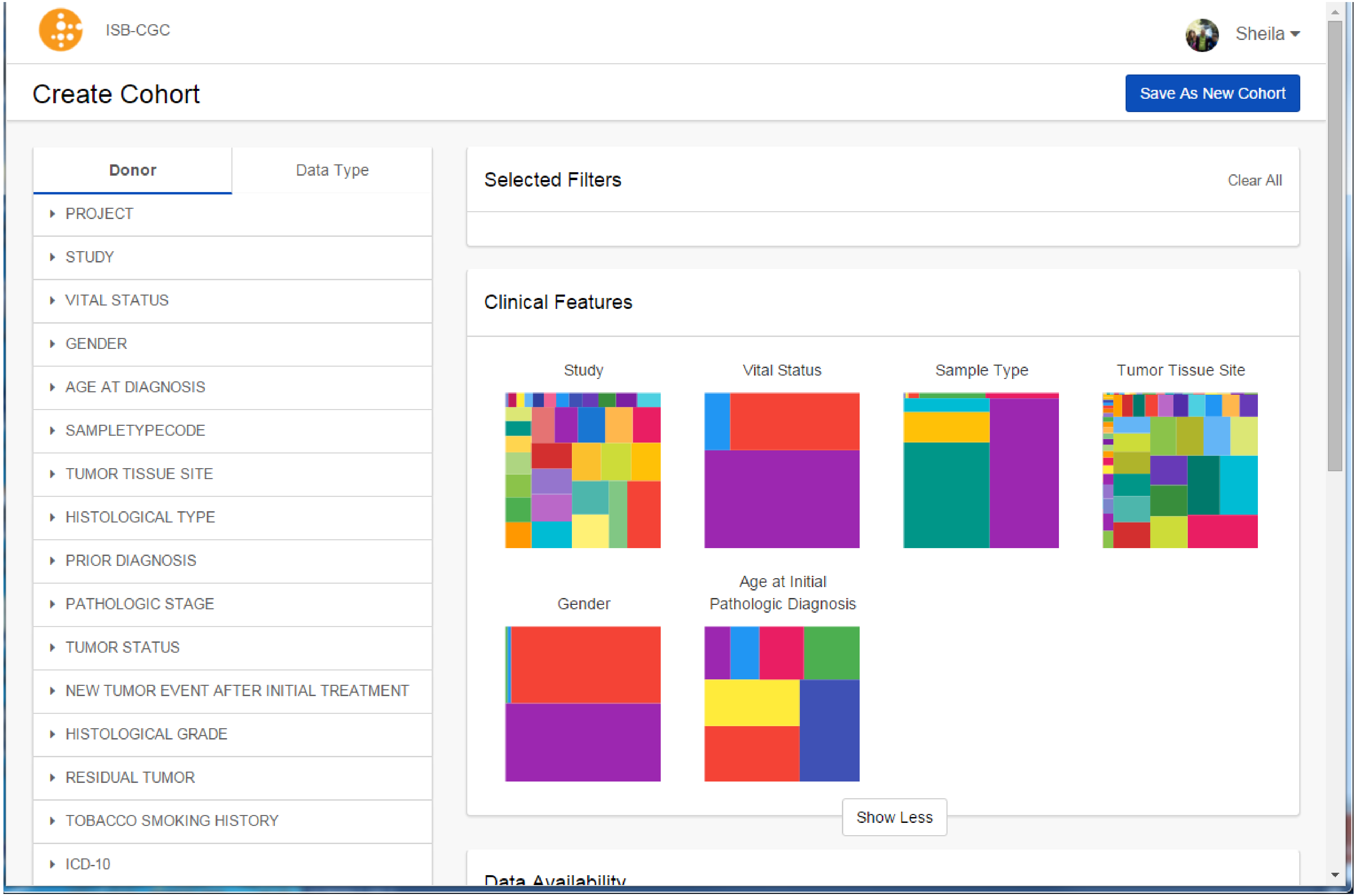
- Code Repository
- Documentation

Warning Notice

You are accessing a U.S. Government (National Institute of Health (NIH), National Cancer Institute (NCI) and Institute for Systems Biology ISB) website / application which may contain information that must be protected under the U. S. Privacy Act or other sensitive information which is intended for authorized Scientific or Research use only. Network activity and Application usage is monitored to maintain system security, availability, and to ensure appropriate and legitimate usage, there is no right or expectation to privacy on this website / system. Unauthorized attempts to upload information, change information, or use of this website for other than its intended purpose may result in disciplinary action, civil, and/or criminal penalties. Any individual or system that accesses this website should have no expectation of privacy regarding any communications or data processed by this website. Anyone accessing this website expressly consents to monitoring of their actions and all communication or data transiting or stored on or related to this website and is advised that if such monitoring reveals possible evidence of criminal activity, NIH may provide that evidence to law enforcement officials.

Funding

This project has been funded in whole with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400007C.



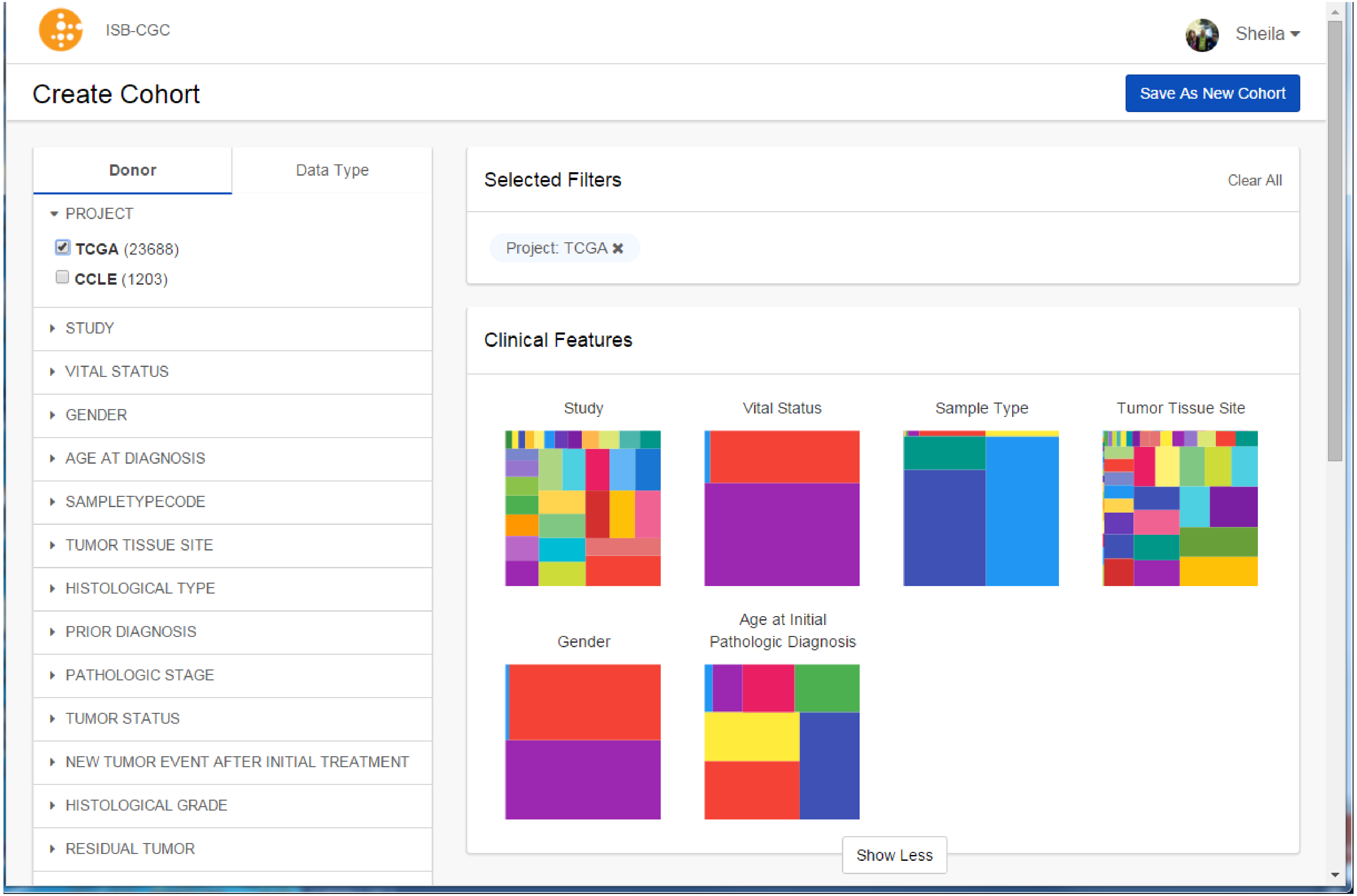
Create Cohort

- PROJECT
- STUDY
- VITAL STATUS
- GENDER
- AGE AT DIAGNOSIS
- SAMPLETYPECODE
- TUMOR TISSUE SITE
- HISTOLOGICAL TYPE
- PRIOR DIAGNOSIS
- PATHOLOGIC STAGE
- TUMOR STATUS
- NEW TUMOR EVENT AFTER INITIAL TREATMENT
- HISTOLOGICAL GRADE
- RESIDUAL TUMOR
- TOBACCO SMOKING HISTORY
- ICD-10

Selected Filters

Clinical Features

Data Availability



Create Cohort

▾ PROJECT

☑ **TCGA** (23688)

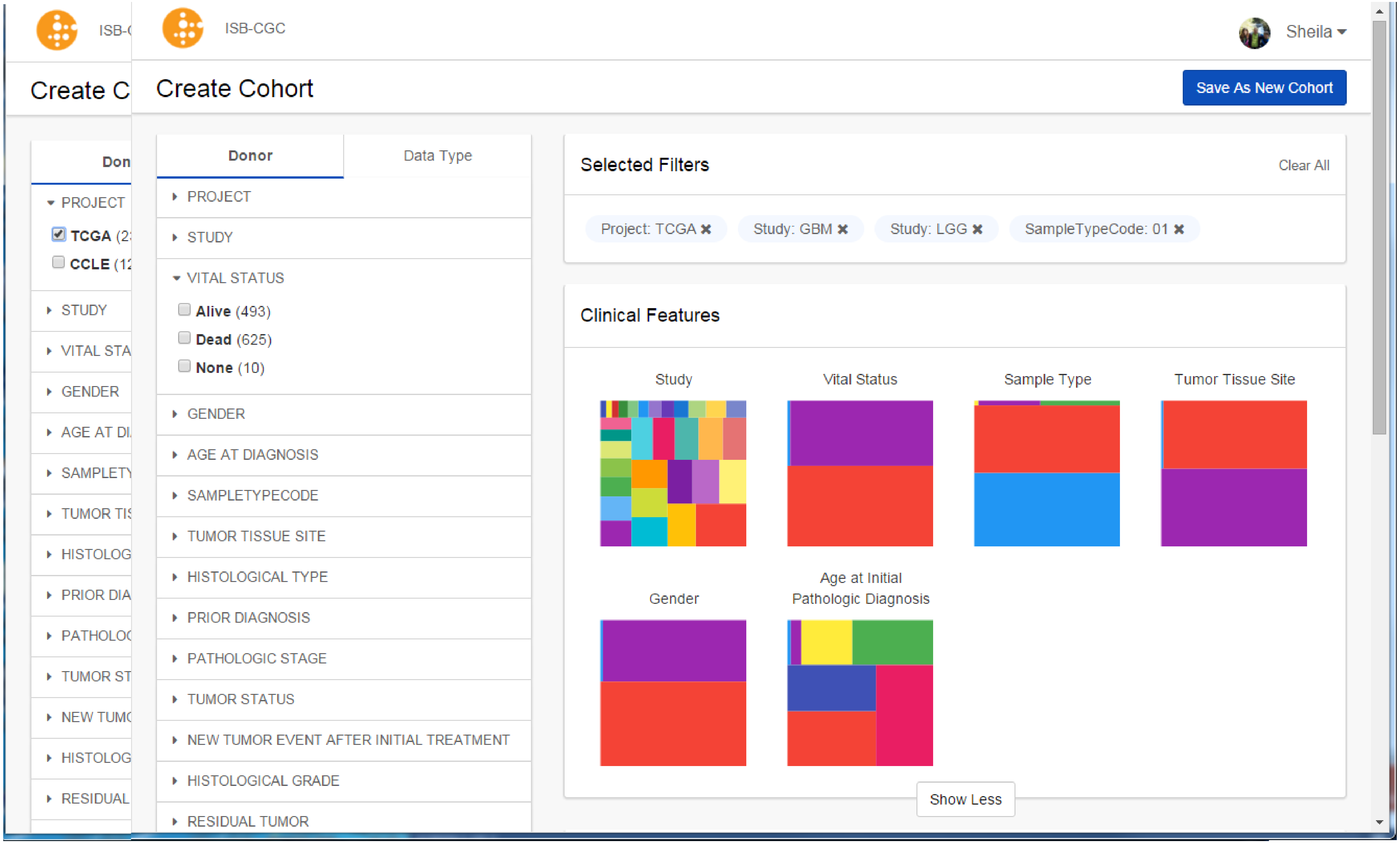
□ **CCLE** (1203)

- ▸ STUDY
- ▸ VITAL STATUS
- ▸ GENDER
- ▸ AGE AT DIAGNOSIS
- ▸ SAMPLETYPECODE
- ▸ TUMOR TISSUE SITE
- ▸ HISTOLOGICAL TYPE
- ▸ PRIOR DIAGNOSIS
- ▸ PATHOLOGIC STAGE
- ▸ TUMOR STATUS
- ▸ NEW TUMOR EVENT AFTER INITIAL TREATMENT
- ▸ HISTOLOGICAL GRADE
- ▸ RESIDUAL TUMOR

Selected Filters

Project: TCGA ✖

Clinical Features

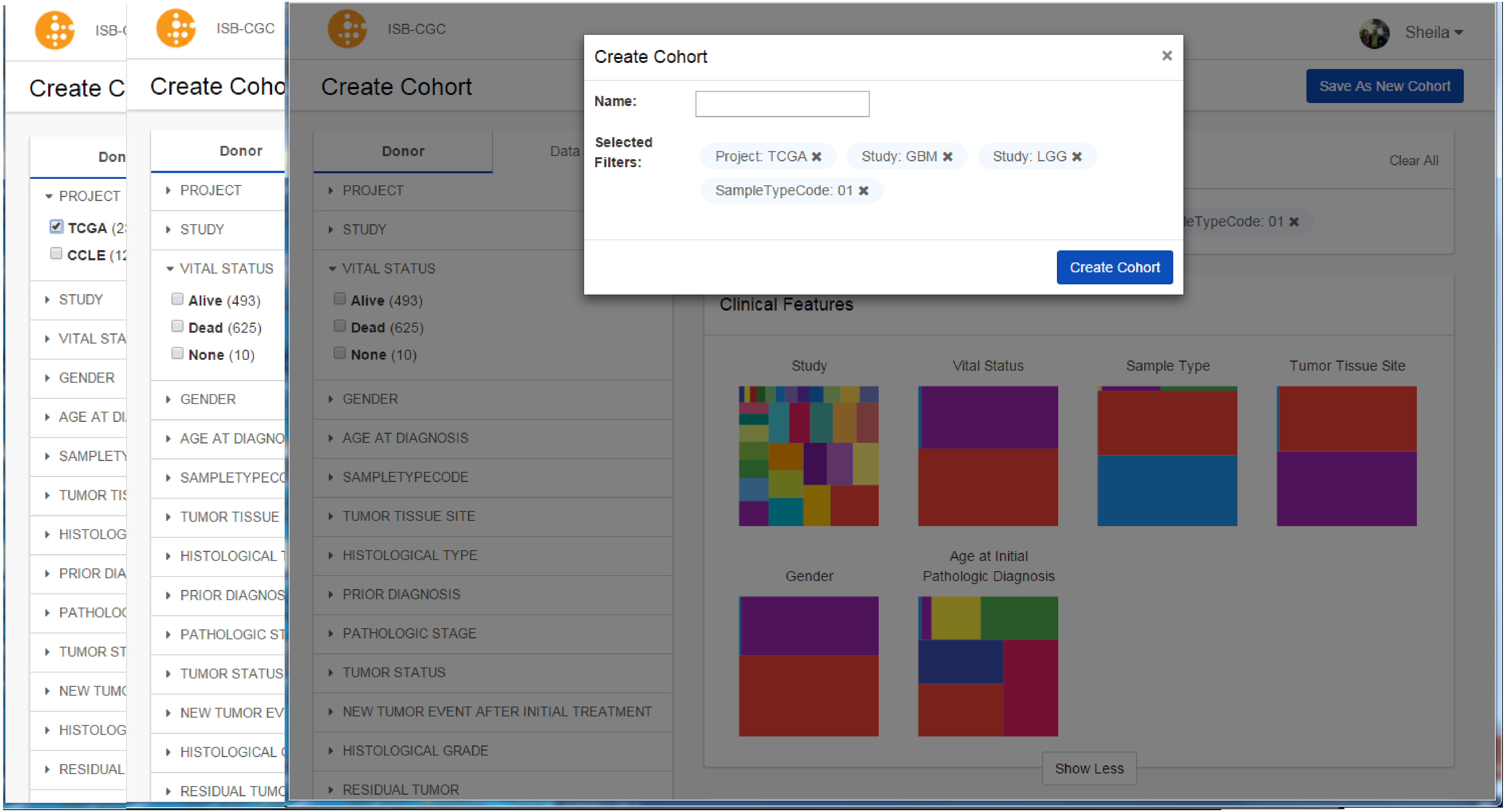


Create Cohort

☐ **Alive** (493)
☐ **Dead** (625)
☐ **None** (10)

Selected Filters

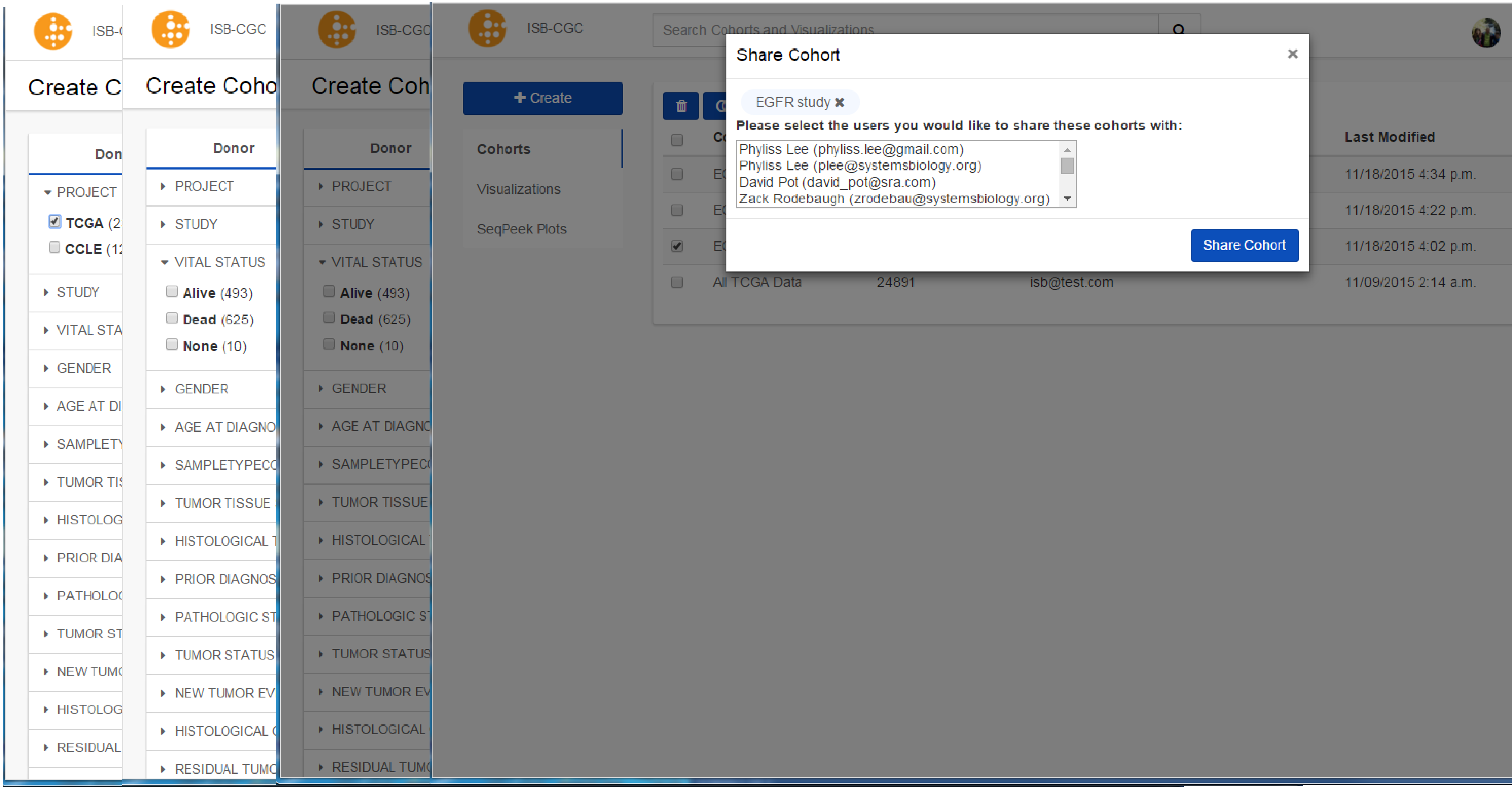
Project: TCGA ✖ Study: GBM ✖ Study: LGG ✖ SampleTypeCode: 01 ✖



Create Cohort

Name:

Selected Filters: Project: TCGA ✖ Study: GBM ✖ Study: LGG ✖ SampleTypeCode: 01 ✖



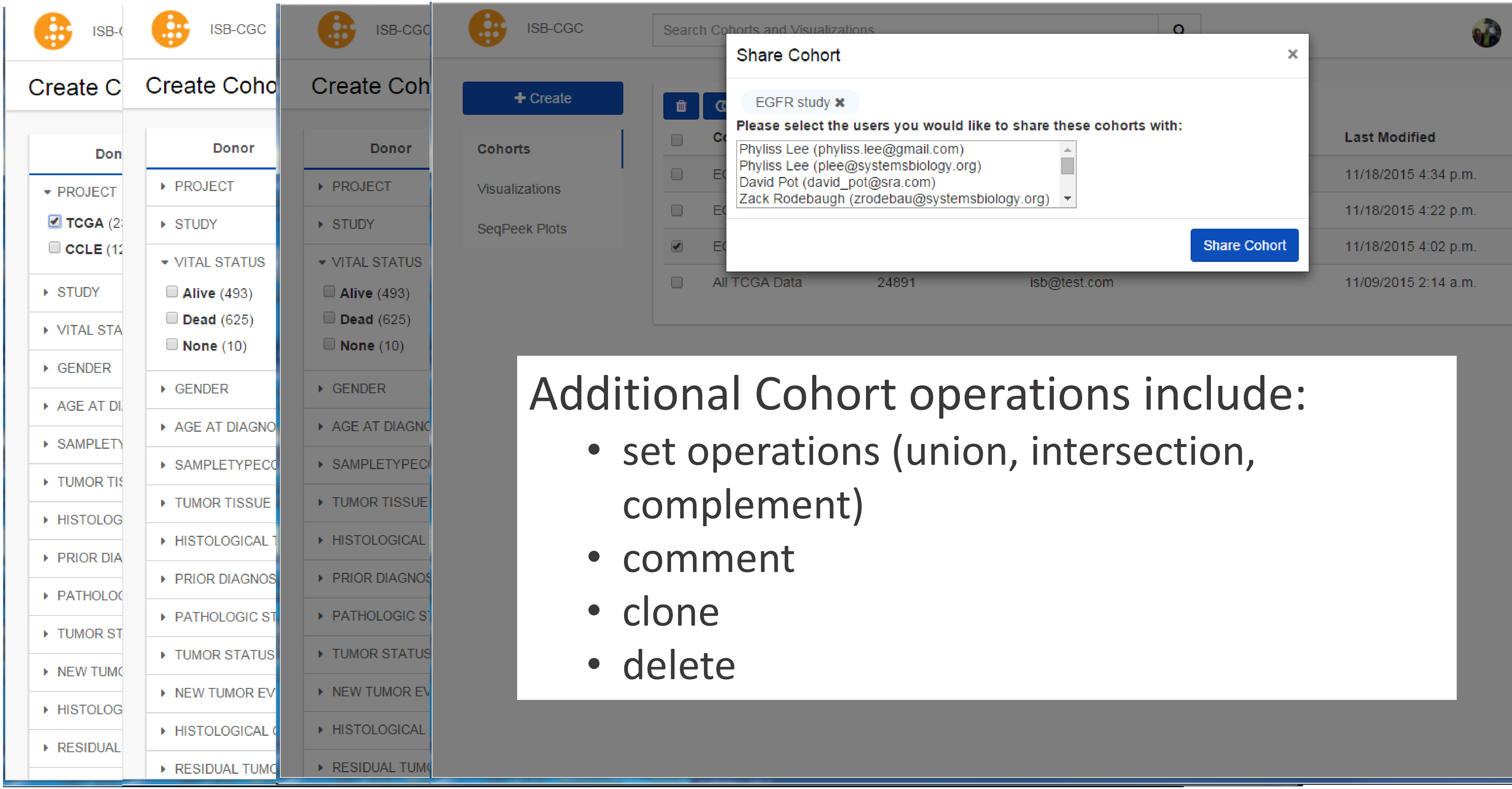
Share Cohort

EGFR study ✖

Please select the users you would like to share these cohorts with:

Phyliss Lee (phyliss.lee@gmail.com)
Phyliss Lee (plee@systemsbiology.org)
David Pot (david_pot@sra.com)
Zack Rodebaugh (zrodebau@systemsbiology.org)

	Co			Last Modified
□	EC			11/18/2015 4:34 p.m.
□	EC			11/18/2015 4:22 p.m.
☑	EC			11/18/2015 4:02 p.m.
□	All TCGA Data	24891	isb@test.com	11/09/2015 2:14 a.m.



Share Cohort

EGFR study ✖

Please select the users you would like to share these cohorts with:

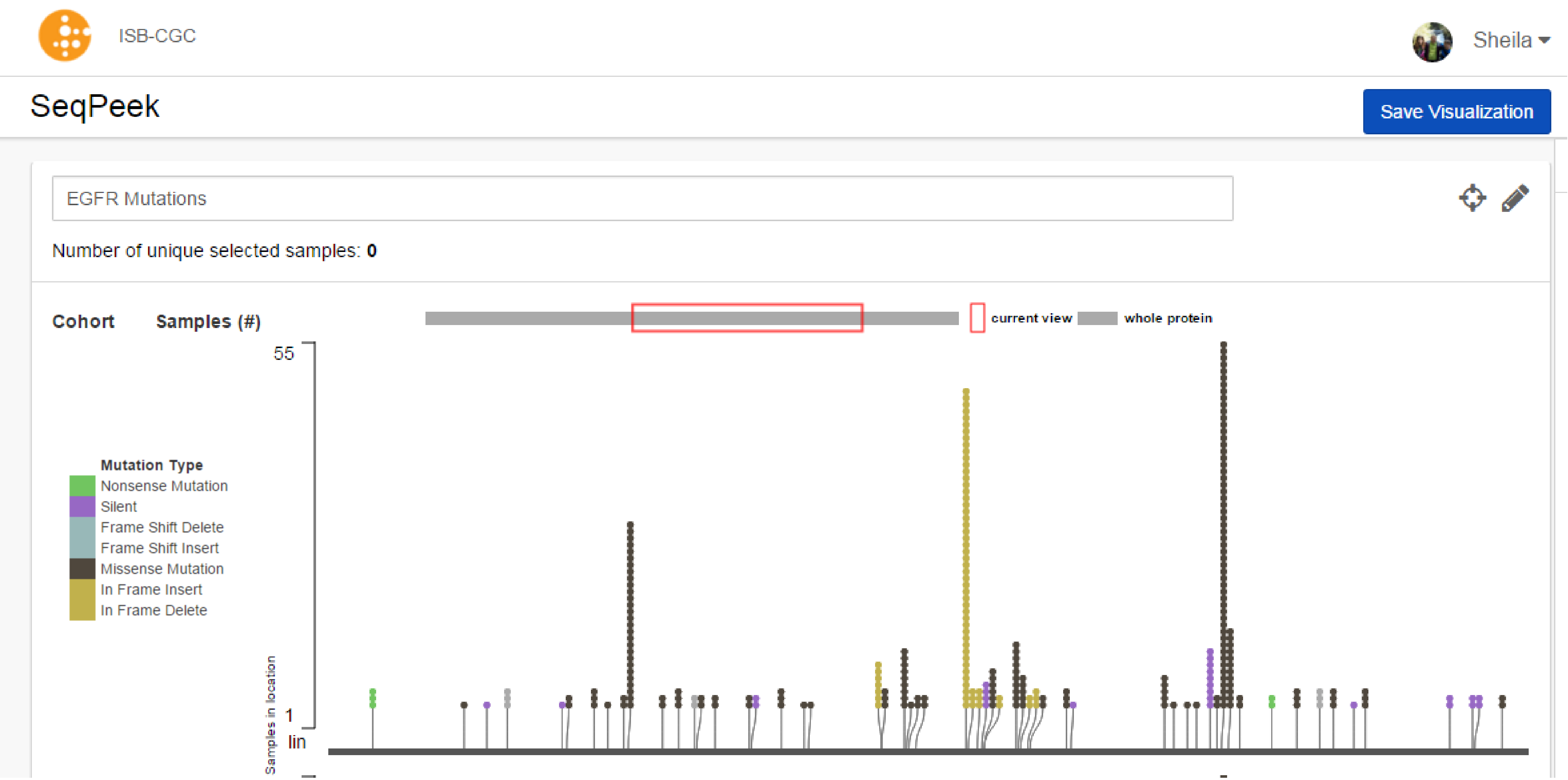
Phyliss Lee (phyliss.lee@gmail.com)
Phyliss Lee (plee@systemsbiology.org)
David Pot (david_pot@sra.com)
Zack Rodebaugh (zrodebau@systemsbiology.org)

Share Cohort

				Last Modified
□	EC			11/18/2015 4:34 p.m.
□	EC			11/18/2015 4:22 p.m.
☑	EC			11/18/2015 4:02 p.m.
□	All TCGA Data	24891	isb@test.com	11/09/2015 2:14 a.m.

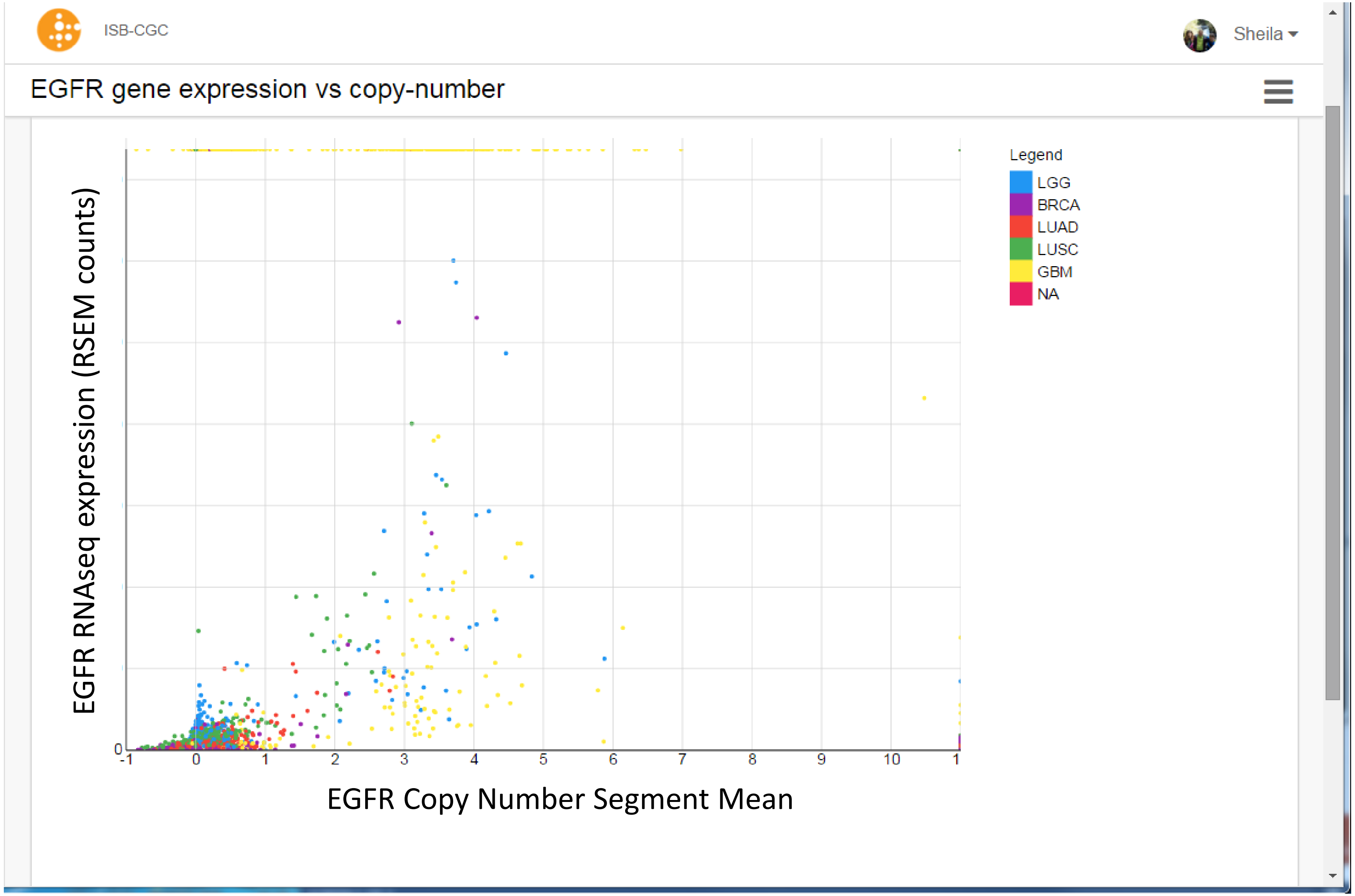
Additional Cohort operations include:

- set operations (union, intersection, complement)
- comment
- clone
- delete



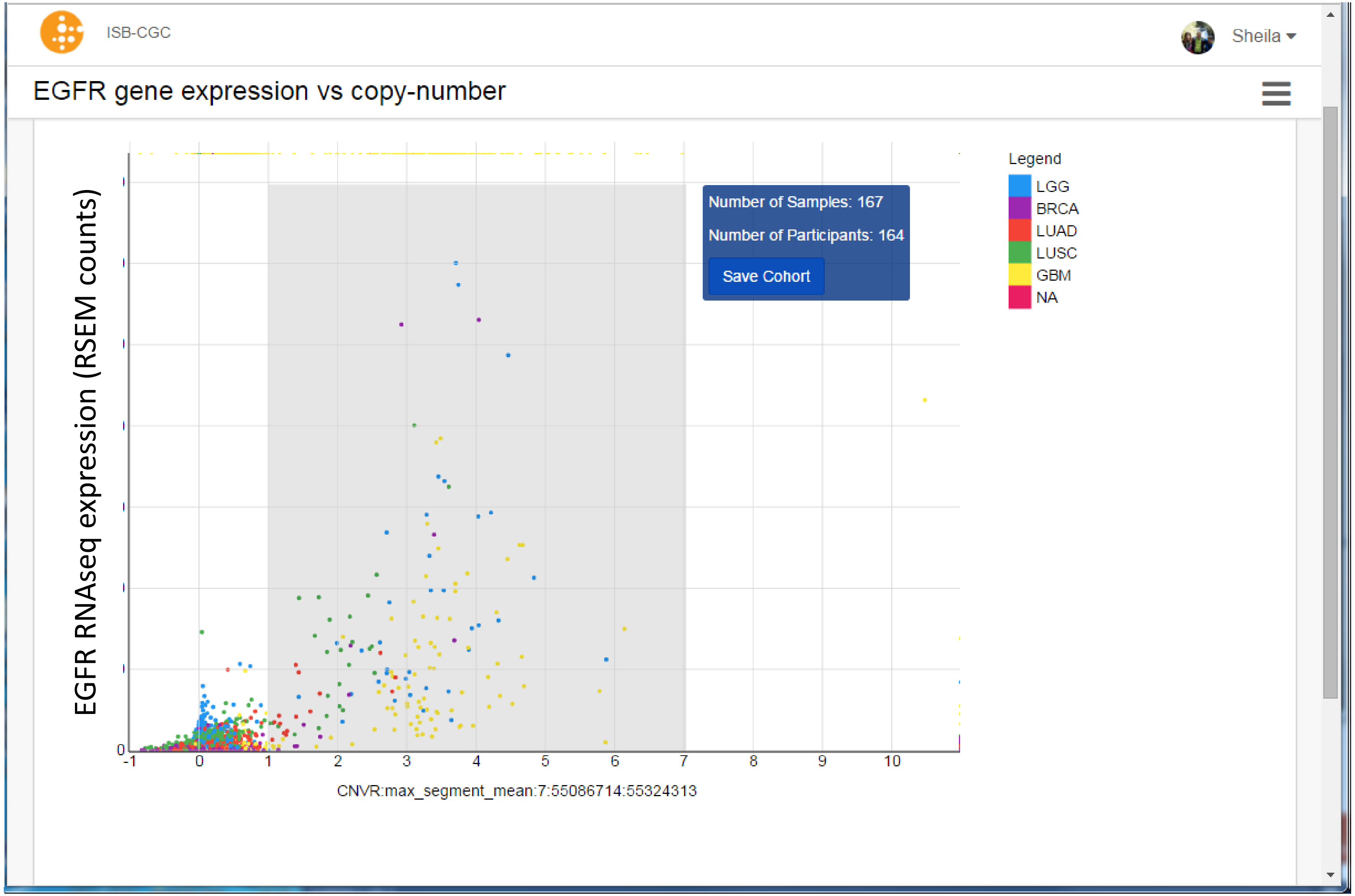
EGFR Mutations

Number of unique selected samples: **0**



ISB-CGC
Sheila
EGFR RNAseq expression (RSEM counts)
0
-1
0
1
2
3
4
5
6
7
8
9
10
1
EGFR Copy Number Segment Mean
Legend
LGG
BRCA
LUAD
LUSC
GBM
NA

EGFR gene expression vs copy-number



ISB-CGC

Sheila

EGFR gene expression vs copy-number

Number of Samples: 167

Number of Participants: 164

Save Cohort

EGFR RNAseq expression (RSEM counts)
0
-1
0
1
2
3
4
5
6
7
8
9
10
CNVR:max_segment_mean:7:55086714:55324313
Legend
LGG
BRCA
LUAD
LUSC
GBM
NA

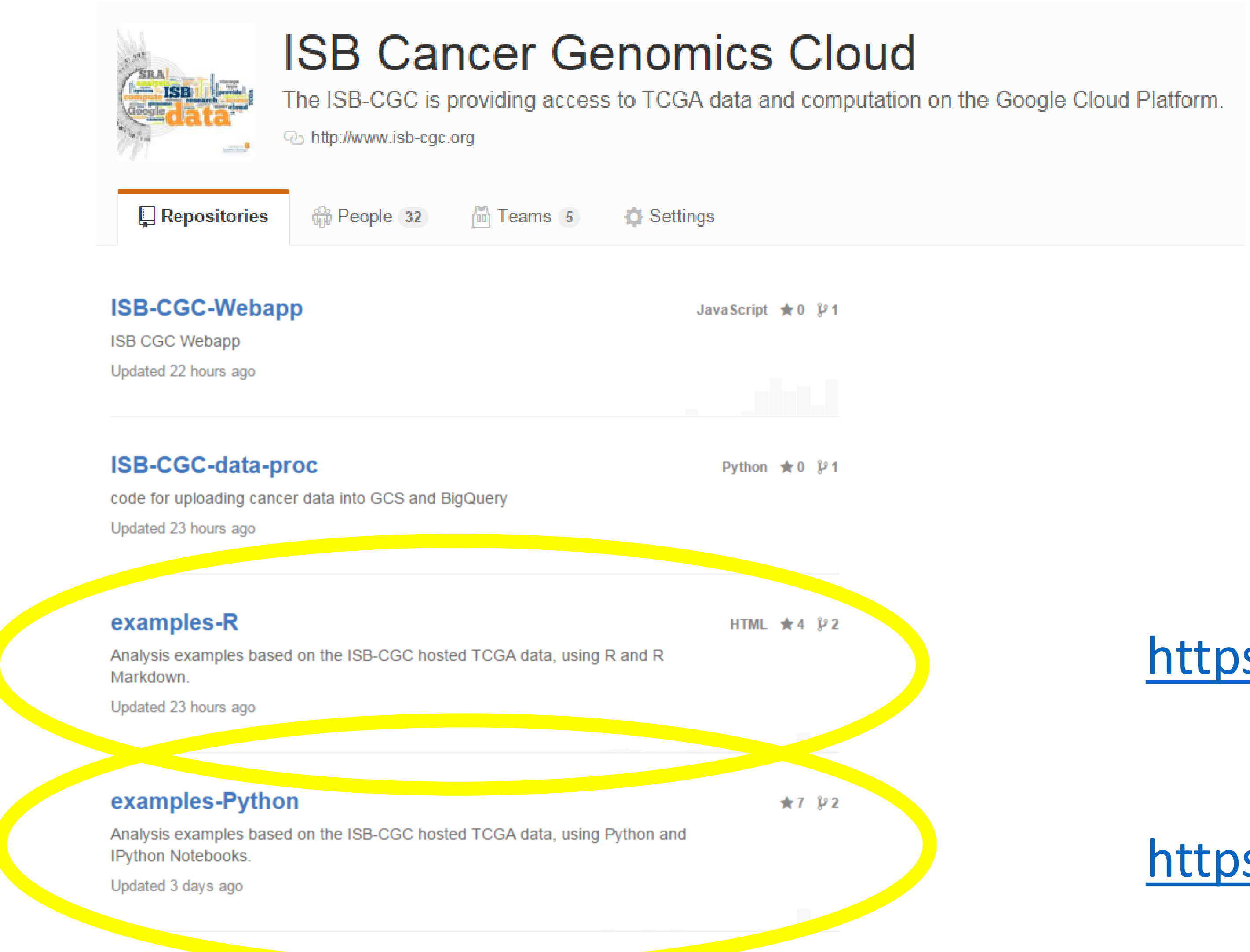
Python, R, and SQL
for the Computational Scientist:

IP[y]: IPython
Interactive Computing

SQL

R Studio®

GitHub



ISB Cancer Genomics Cloud

The ISB-CGC is providing access to TCGA data and computation on the Google Cloud Platform.

http://www.isb-cgc.org

ISB-CGC-Webapp JavaScript ★ 0 ⑂ 1

ISB CGC Webapp

Updated 22 hours ago

ISB-CGC-data-proc Python ★ 0 ⑂ 1

code for uploading cancer data into GCS and BigQuery

Updated 23 hours ago

examples-R HTML ★ 4 ⑂ 2

Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

Updated 23 hours ago

examples-Python ★ 7 ⑂ 2

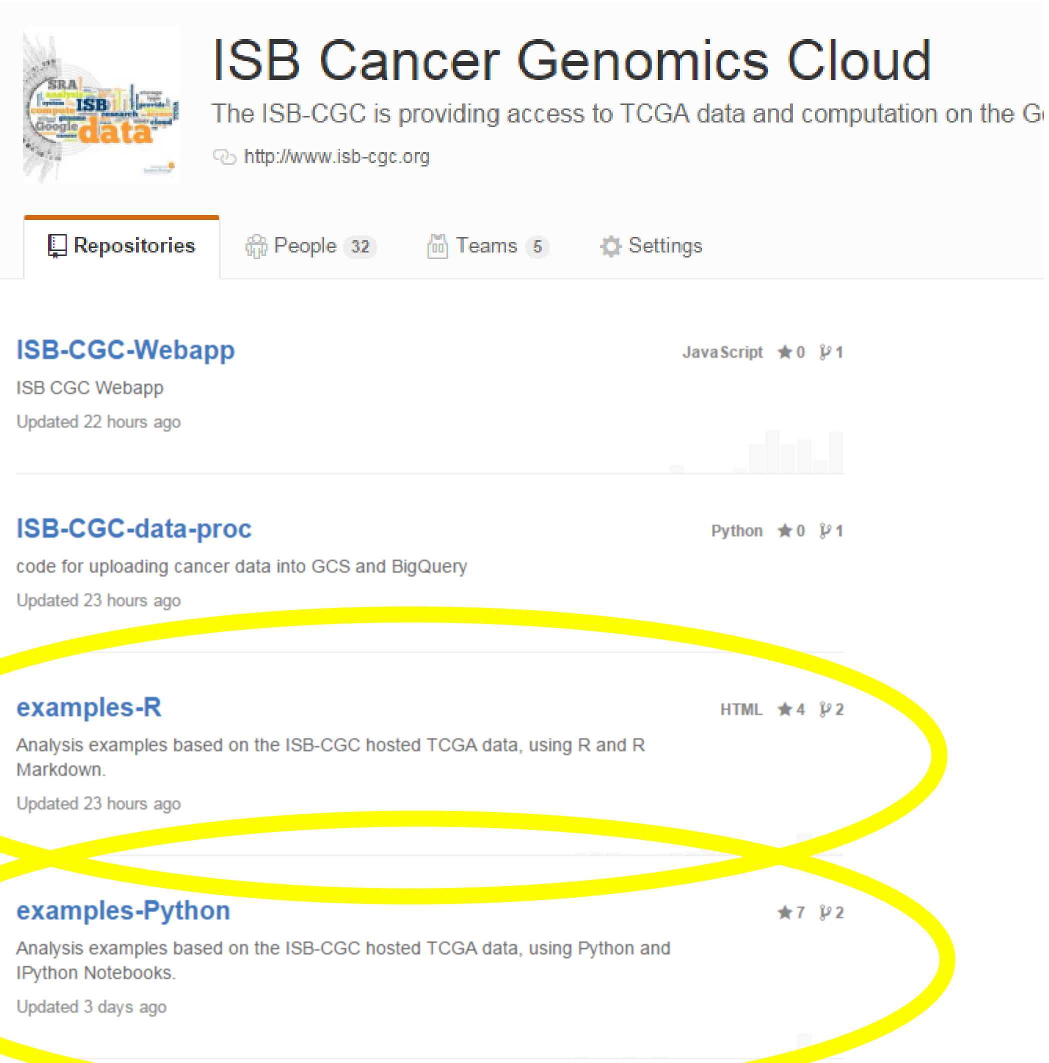
Analysis examples based on the ISB-CGC hosted TCGA data, using Python and IPython Notebooks.

Updated 3 days ago

https://github.com/isb-cgc/examples-R

https://github.com/isb-cgc/examples-Python

GitHub



ISB Cancer Genomics Cloud

The ISB-CGC is providing access to TCGA data and computation on the G

http://www.isb-cgc.org

ISB-CGC-Webapp JavaScript ★0 1

ISB CGC Webapp

Updated 22 hours ago

ISB-CGC-data-proc Python ★0 1

code for uploading cancer data into GCS and BigQuery

Updated 23 hours ago

examples-R HTML ★4 2

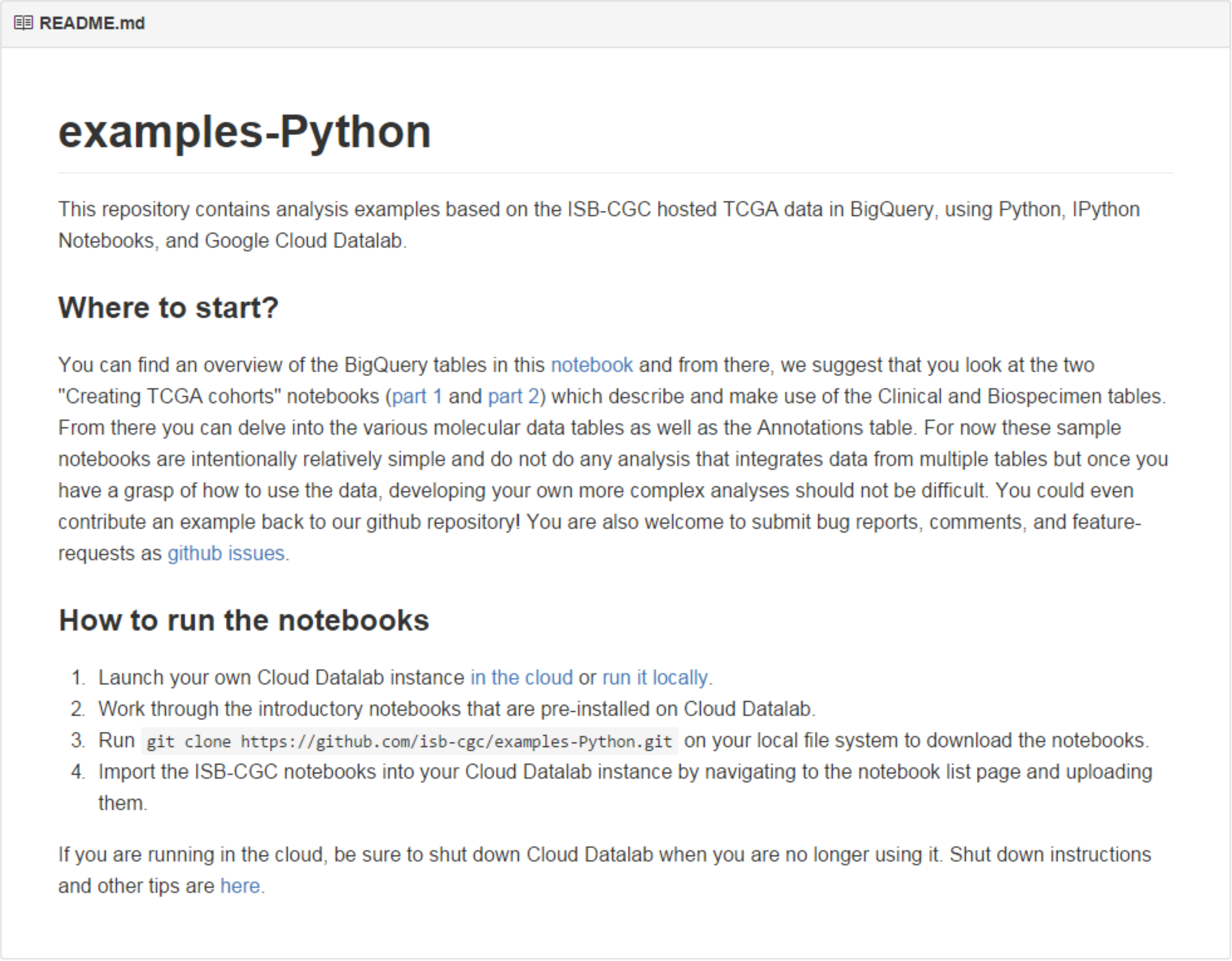
Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

Updated 23 hours ago

examples-Python ★7 2

Analysis examples based on the ISB-CGC hosted TCGA data, using Python and IPython Notebooks.

Updated 3 days ago



README.md

examples-Python

This repository contains analysis examples based on the ISB-CGC hosted TCGA data in BigQuery, using Python, IPython Notebooks, and Google Cloud Datalab.

Where to start?

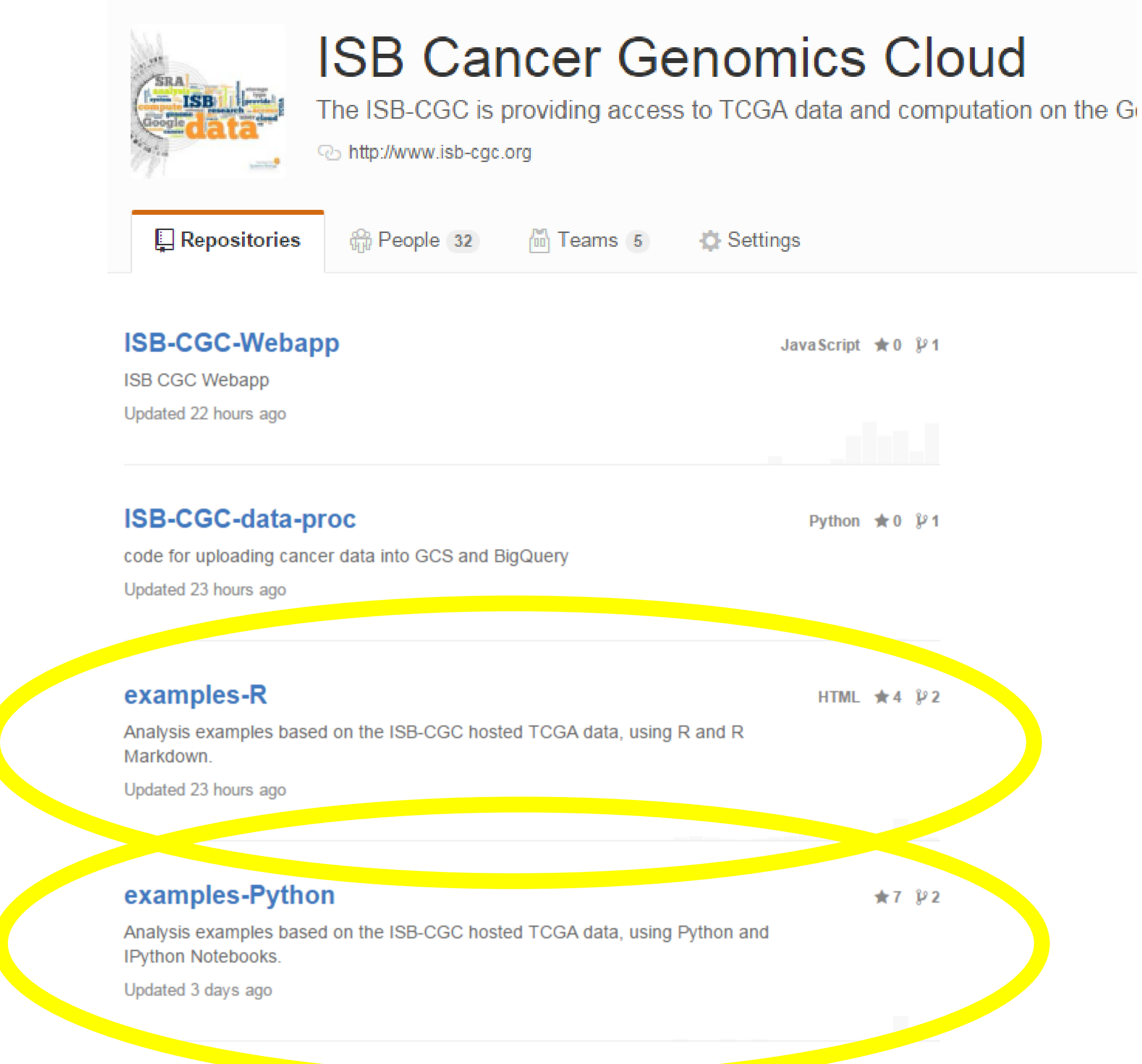
You can find an overview of the BigQuery tables in this notebook and from there, we suggest that you look at the two "Creating TCGA cohorts" notebooks (part 1 and part 2) which describe and make use of the Clinical and Biospecimen tables. From there you can delve into the various molecular data tables as well as the Annotations table. For now these sample notebooks are intentionally relatively simple and do not do any analysis that integrates data from multiple tables but once you have a grasp of how to use the data, developing your own more complex analyses should not be difficult. You could even contribute an example back to our github repository! You are also welcome to submit bug reports, comments, and feature-requests as github issues.

How to run the notebooks

1. Launch your own Cloud Datalab instance in the cloud or run it locally.
2. Work through the introductory notebooks that are pre-installed on Cloud Datalab.
3. Run `git clone https://github.com/isb-cgc/examples-Python.git` on your local file system to download the notebooks.
4. Import the ISB-CGC notebooks into your Cloud Datalab instance by navigating to the notebook list page and uploading them.

If you are running in the cloud, be sure to shut down Cloud Datalab when you are no longer using it. Shut down instructions and other tips are here.

GitHub



ISB Cancer Genomics Cloud

The ISB-CGC is providing access to TCGA data and computation on the G

http://www.isb-cgc.org

ISB-CGC-Webapp JavaScript ★0 1

ISB CGC Webapp

Updated 22 hours ago

ISB-CGC-data-proc Python ★0 1

code for uploading cancer data into GCS and BigQuery

Updated 23 hours ago

examples-R HTML ★4 2

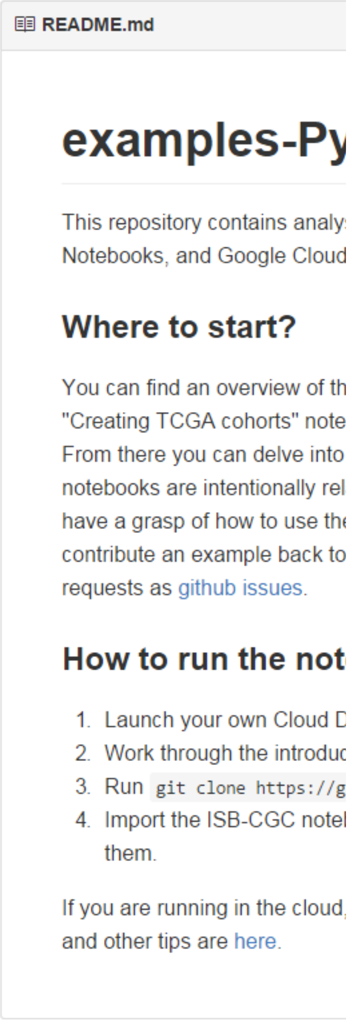
Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

Updated 23 hours ago

examples-Python ★7 2

Analysis examples based on the ISB-CGC hosted TCGA data, using Python and IPython Notebooks.

Updated 3 days ago



README.md

examples-Py

This repository contains analy Notebooks, and Google Cloud

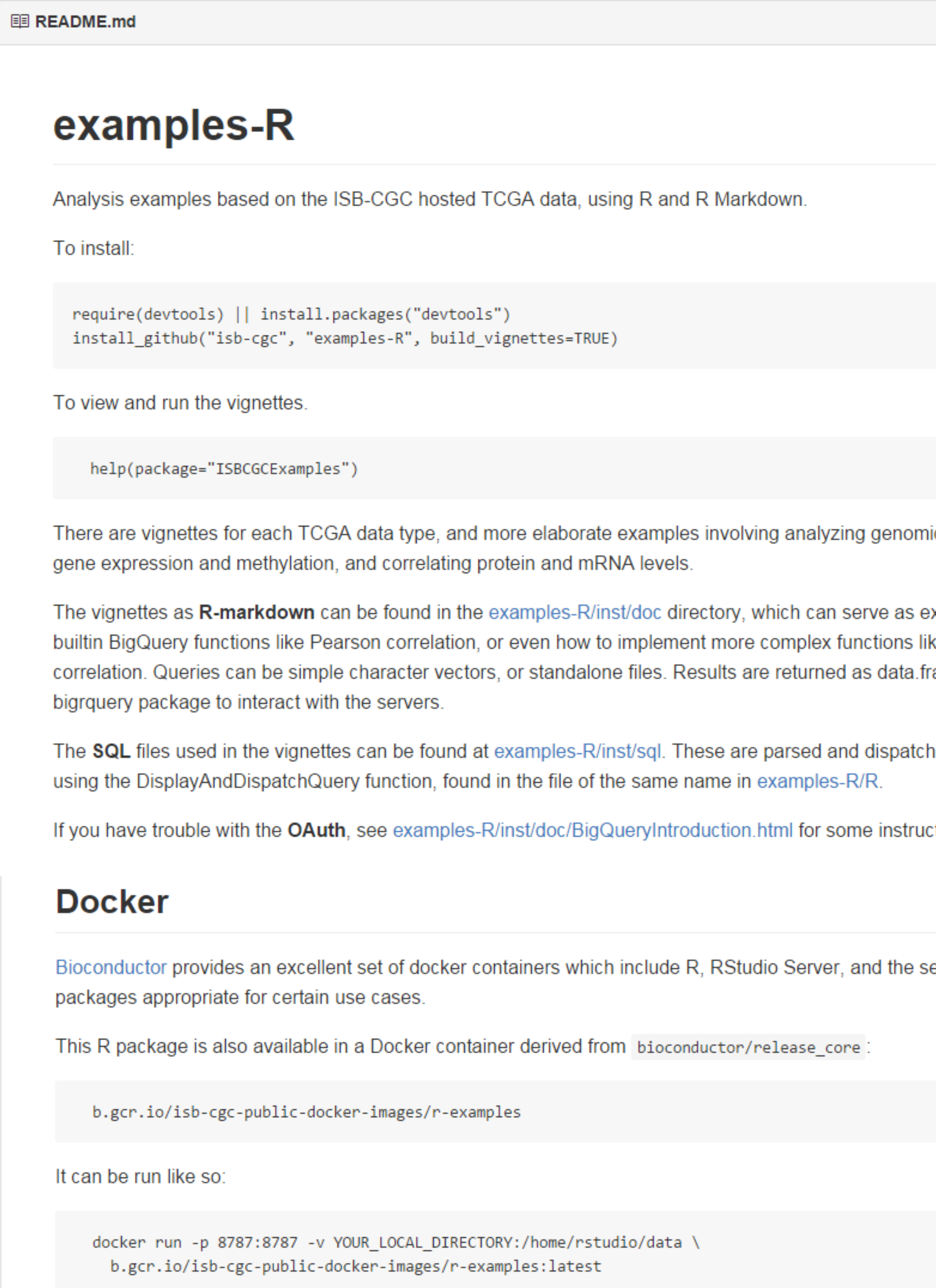
Where to start?

You can find an overview of th "Creating TCGA cohorts" note From there you can delve into notebooks are intentionally rel have a grasp of how to use the contribute an example back to requests as github issues.

How to run the not

1. Launch your own Cloud D
2. Work through the introduc
3. Run `git clone https://g`
4. Import the ISB-CGC notel them.

If you are running in the cloud, and other tips are here.



README.md

examples-R

Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

To install:

```
require(devtools) || install.packages("devtools")
install_github("isb-cgc", "examples-R", build_vignettes=TRUE)
```

To view and run the vignettes.

```
help(package="ISBCGCExamples")
```

There are vignettes for each TCGA data type, and more elaborate examples involving analyzing genomi gene expression and methylation, and correlating protein and mRNA levels.

The vignettes as **R-markdown** can be found in the examples-R/inst/doc directory, which can serve as ex builtin BigQuery functions like Pearson correlation, or even how to implement more complex functions lik correlation. Queries can be simple character vectors, or standalone files. Results are returned as data.fra bigrquery package to interact with the servers.

The **SQL** files used in the vignettes can be found at examples-R/inst/sql. These are parsed and dispatch using the DisplayAndDispatchQuery function, found in the file of the same name in examples-R/R.

If you have trouble with the **OAuth**, see examples-R/inst/doc/BigQueryIntroduction.html for some instruc

Docker

Bioconductor provides an excellent set of docker containers which include R, RStudio Server, and the se packages appropriate for certain use cases.

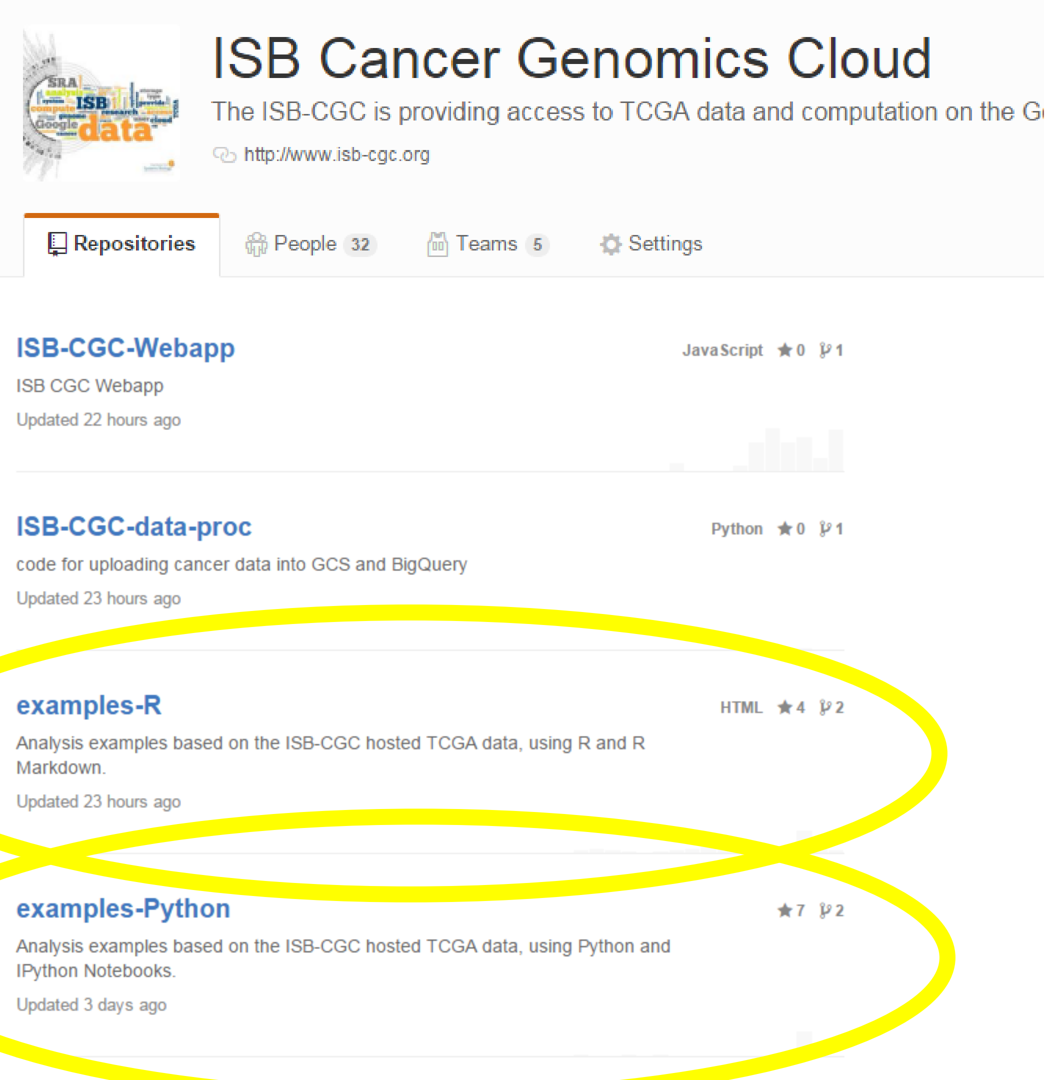
This R package is also available in a Docker container derived from `bioconductor/release_core`:

```
b.gcr.io/isb-cgc-public-docker-images/r-examples
```

It can be run like so:

```
docker run -p 8787:8787 -v YOUR_LOCAL_DIRECTORY:/home/rstudio/data \
  b.gcr.io/isb-cgc-public-docker-images/r-examples:latest
```

GitHub



ISB Cancer Genomics Cloud

The ISB-CGC is providing access to TCGA data and computation on the G

http://www.isb-cgc.org

ISB-CGC-Webapp — JavaScript ★0 1

ISB CGC Webapp

Updated 22 hours ago

ISB-CGC-data-proc — Python ★0 1

code for uploading cancer data into GCS and BigQuery

Updated 23 hours ago

examples-R — HTML ★4 2

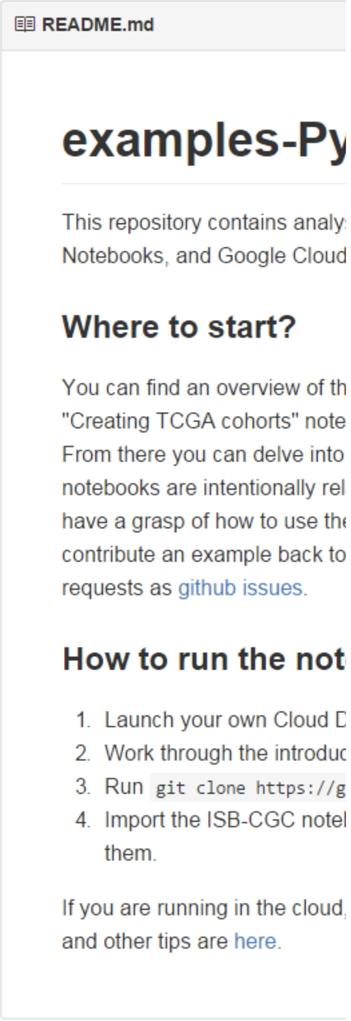
Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

Updated 23 hours ago

examples-Python — ★7 2

Analysis examples based on the ISB-CGC hosted TCGA data, using Python and IPython Notebooks.

Updated 3 days ago



README.md

examples-Py

This repository contains analy Notebooks, and Google Cloud

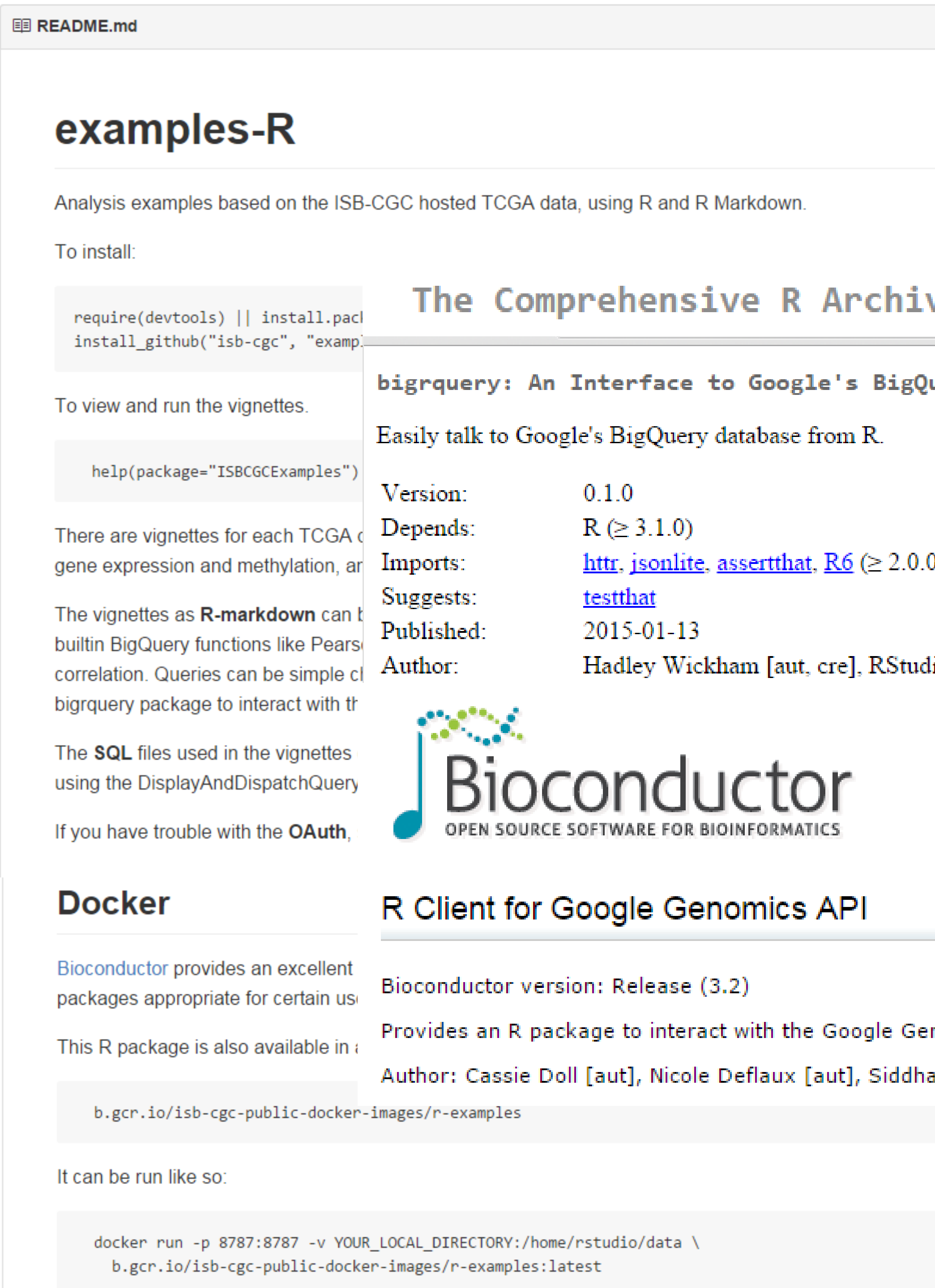
Where to start?

You can find an overview of th "Creating TCGA cohorts" note From there you can delve into notebooks are intentionally rel have a grasp of how to use the contribute an example back to requests as github issues.

How to run the not

1. Launch your own Cloud D
2. Work through the introduc
3. Run `git clone https://g`
4. Import the ISB-CGC notel them.

If you are running in the cloud, and other tips are here.



README.md

examples-R

Analysis examples based on the ISB-CGC hosted TCGA data, using R and R Markdown.

To install:

```
require(devtools) || install.pacl
install_github("isb-cgc", "examp
```

To view and run the vignettes.

```
help(package="ISBCGCExamples")
```

There are vignettes for each TCGA c gene expression and methylation, ar

The vignettes as **R-markdown** can b builtin BigQuery functions like Pears correlation. Queries can be simple cl bigrquery package to interact with th

The **SQL** files used in the vignettes using the DisplayAndDispatchQuery

If you have trouble with the **OAuth**,

Docker

Bioconductor provides an excellent packages appropriate for certain us

This R package is also available in a

```
b.gcr.io/isb-cgc-public-docker-images/r-examples
```

It can be run like so:

```
docker run -p 8787:8787 -v YOUR_LOCAL_DIRECTORY:/home/rstudio/data \
  b.gcr.io/isb-cgc-public-docker-images/r-examples:latest
```

The Comprehensive R Archiv

bigrquery: An Interface to Google's BigQu

Easily talk to Google's BigQuery database from R.

Version:	0.1.0
Depends:	R (≥ 3.1.0)
Imports:	httr, jsonlite, assertthat, R6 (≥ 2.0.0
Suggests:	testthat
Published:	2015-01-13
Author:	Hadley Wickham [aut, cre], RStudi

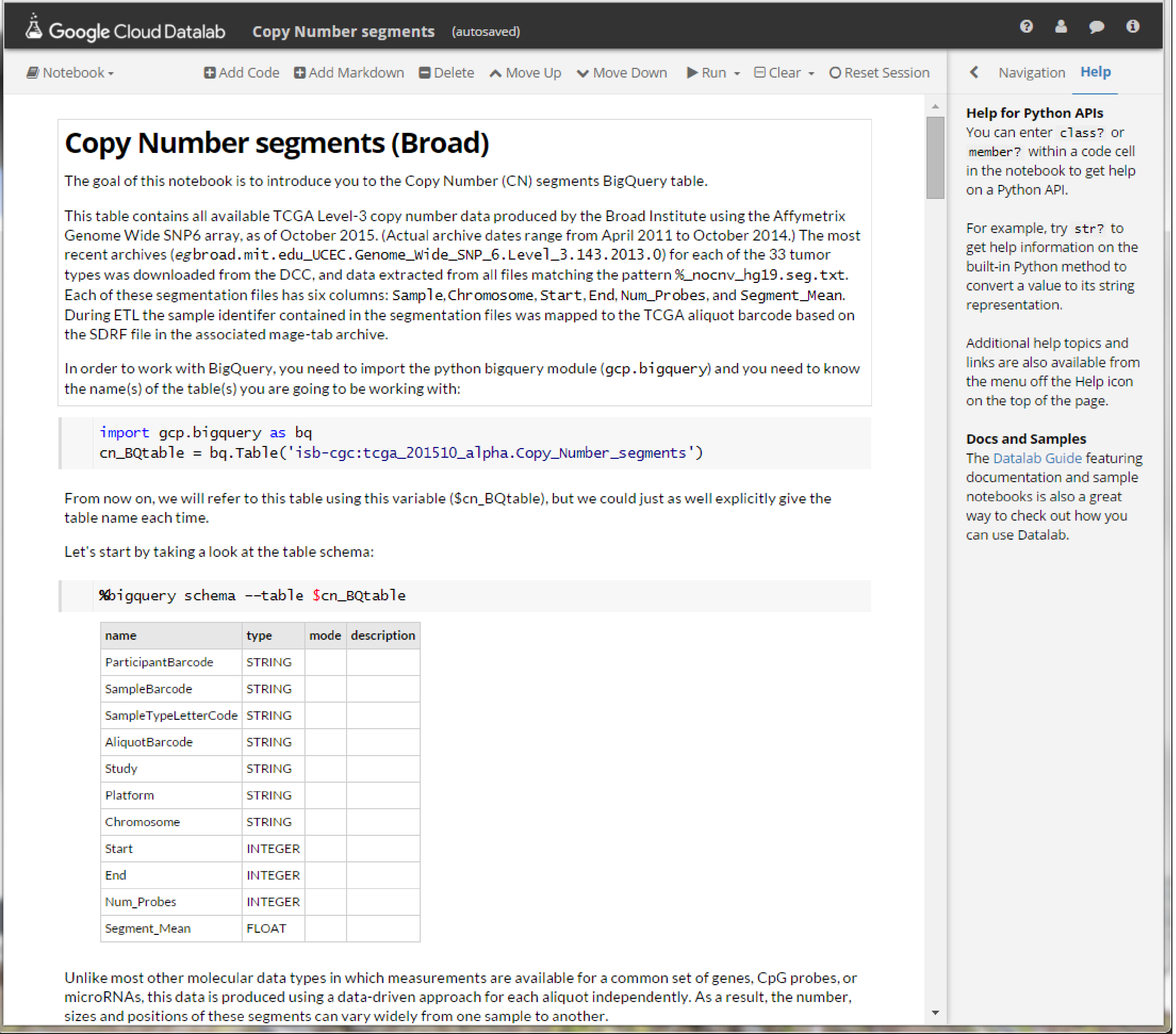
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

R Client for Google Genomics API

Bioconductor version: Release (3.2)

Provides an R package to interact with the Google Ger

Author: Cassie Doll [aut], Nicole Deflaux [aut], Siddha



Copy Number segments (Broad)

The goal of this notebook is to introduce you to the Copy Number (CN) segments BigQuery table.

This table contains all available TCGA Level-3 copy number data produced by the Broad Institute using the Affymetrix Genome Wide SNP6 array, as of October 2015. (Actual archive dates range from April 2011 to October 2014.) The most recent archives (*eg*`broad.mit.edu_UCEC.Genome_Wide_SNP_6.Level_3.143.2013.0`) for each of the 33 tumor types was downloaded from the DCC, and data extracted from all files matching the pattern `%_nocnv_hg19.seg.txt`. Each of these segmentation files has six columns: `Sample`, `Chromosome`, `Start`, `End`, `Num_Probes`, and `Segment_Mean`. During ETL the sample identifer contained in the segmentation files was mapped to the TCGA aliquot barcode based on the SDRF file in the associated mage-tab archive.

In order to work with BigQuery, you need to import the python bigquery module (`gcp.bigquery`) and you need to know the name(s) of the table(s) you are going to be working with:

```
import gcp.bigquery as bq
cn_BQtable = bq.Table('isb-cgc:tcga_201510_alpha.Copy_Number_segments')
```

From now on, we will refer to this table using this variable ($cn_BQtable), but we could just as well explicitly give the table name each time.

Let's start by taking a look at the table schema:

```
%bigquery schema --table $cn_BQtable
```

name	type	mode	description
ParticipantBarcode	STRING		
SampleBarcode	STRING		
SampleTypeLetterCode	STRING		
AliquotBarcode	STRING		
Study	STRING		
Platform	STRING		
Chromosome	STRING		
Start	INTEGER		
End	INTEGER		
Num_Probes	INTEGER		
Segment_Mean	FLOAT		

Unlike most other molecular data types in which measurements are available for a common set of genes, CpG probes, or microRNAs, this data is produced using a data-driven approach for each aliquot independently. As a result, the number, sizes and positions of these segments can vary widely from one sample to another.

Help for Python APIs

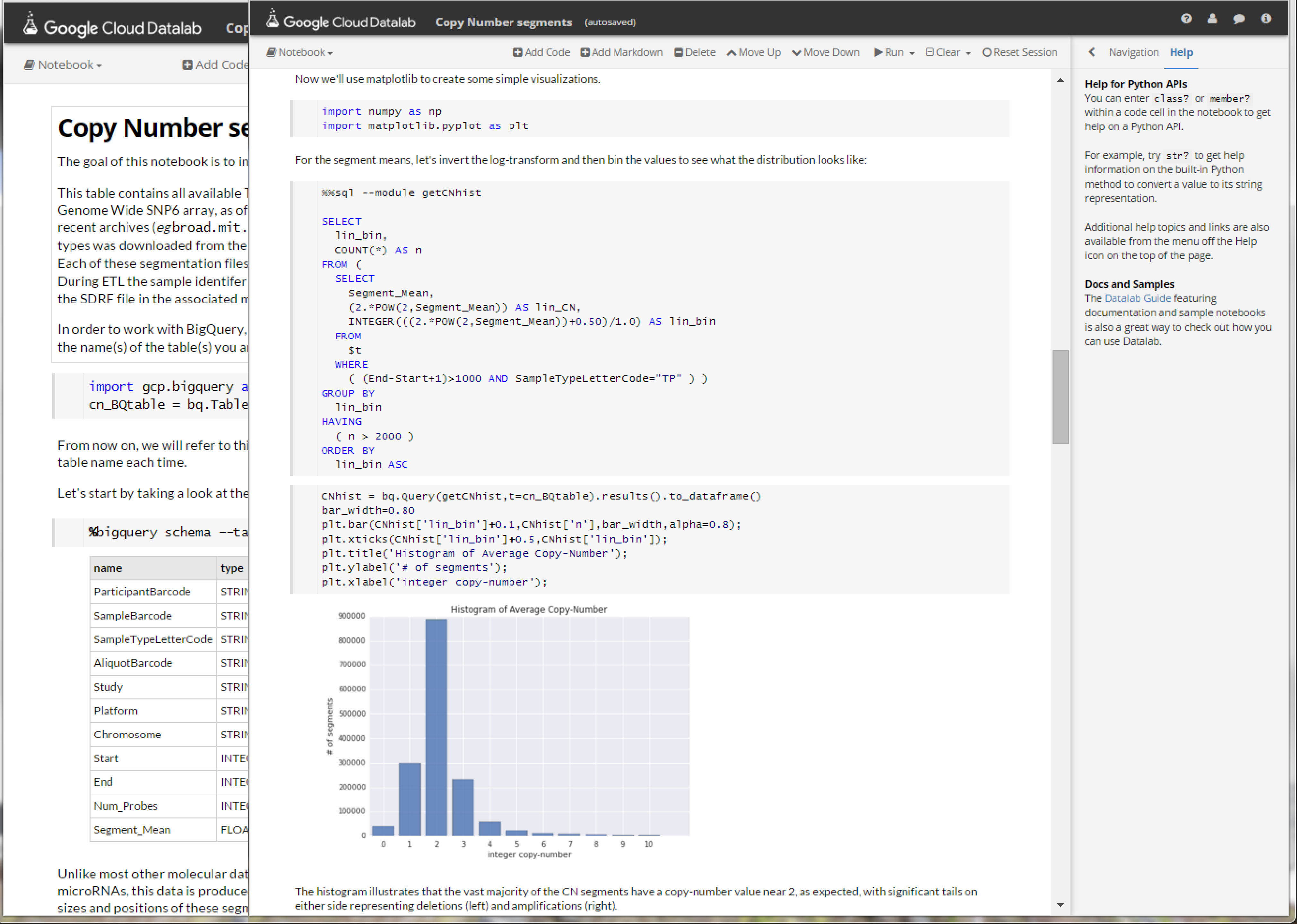
You can enter `class?` or `member?` within a code cell in the notebook to get help on a Python API.

For example, try `str?` to get help information on the built-in Python method to convert a value to its string representation.

Additional help topics and links are also available from the menu off the Help icon on the top of the page.

Docs and Samples

The Datalab Guide featuring documentation and sample notebooks is also a great way to check out how you can use Datalab.



Copy Number se

The goal of this notebook is to in

This table contains all available T
Genome Wide SNP6 array, as of
recent archives (*eg*broad.mit.
types was downloaded from the
Each of these segmentation files
During ETL the sample identifer
the SDRF file in the associated m

In order to work with BigQuery,
the name(s) of the table(s) you a

```
import gcp.bigquery a
cn_BQtable = bq.Table
```

From now on, we will refer to thi
table name each time.

Let's start by taking a look at the

```
%bigquery schema --ta
```

name	type
ParticipantBarcode	STRIN
SampleBarcode	STRIN
SampleTypeLetterCode	STRIN
AliquotBarcode	STRIN
Study	STRIN
Platform	STRIN
Chromosome	STRIN
Start	INTE
End	INTE
Num_Probes	INTE
Segment_Mean	FLOA

Unlike most other molecular dat
microRNAs, this data is produce
sizes and positions of these segm

Now we'll use matplotlib to create some simple visualizations.

```
import numpy as np
import matplotlib.pyplot as plt
```

For the segment means, let's invert the log-transform and then bin the values to see what the distribution looks like:

```
%%sql --module getCNhist

SELECT
  lin_bin,
  COUNT(*) AS n
FROM (
  SELECT
    Segment_Mean,
    (2.*POW(2,Segment_Mean)) AS lin_CN,
    INTEGER(((2.*POW(2,Segment_Mean))+0.50)/1.0) AS lin_bin
  FROM
    $t
  WHERE
    ( (End-Start+1)>1000 AND SampleTypeLetterCode="TP" ) )
GROUP BY
  lin_bin
HAVING
  ( n > 2000 )
ORDER BY
  lin_bin ASC
```

```
CNhist = bq.Query(getCNhist,t=cn_BQtable).results().to_dataframe()
bar_width=0.80
plt.bar(CNhist['lin_bin']+0.1,CNhist['n'],bar_width,alpha=0.8);
plt.xticks(CNhist['lin_bin']+0.5,CNhist['lin_bin']);
plt.title('Histogram of Average Copy-Number');
plt.ylabel('# of segments');
plt.xlabel('integer copy-number');
```

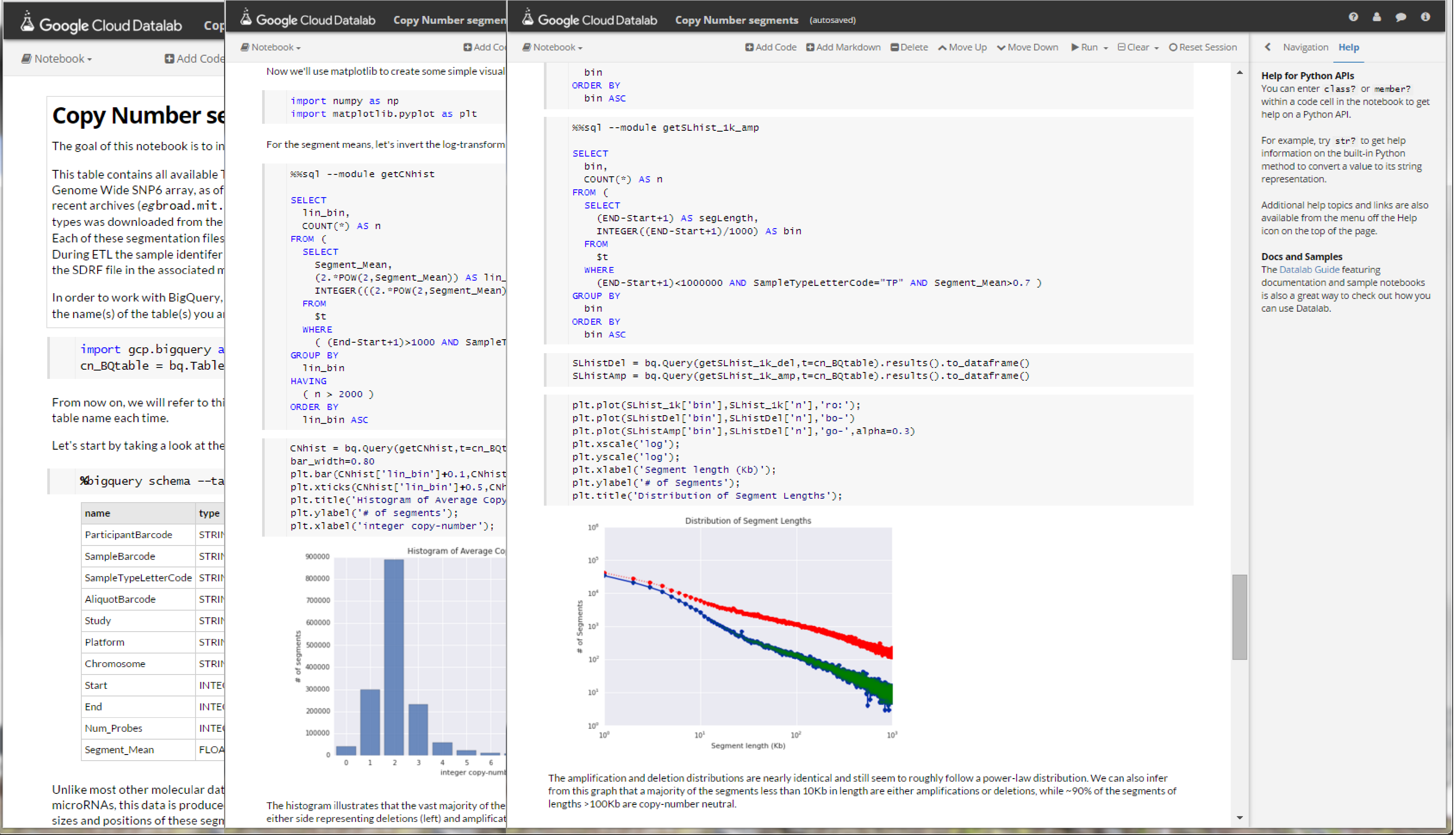
The histogram illustrates that the vast majority of the CN segments have a copy-number value near 2, as expected, with significant tails on either side representing deletions (left) and amplifications (right).

Help for Python APIs
You can enter `class?` or `member?` within a code cell in the notebook to get help on a Python API.

For example, try `str?` to get help information on the built-in Python method to convert a value to its string representation.

Additional help topics and links are also available from the menu off the Help icon on the top of the page.

Docs and Samples
The Datalab Guide featuring documentation and sample notebooks is also a great way to check out how you can use Datalab.



Copy Number se

The goal of this notebook is to in

This table contains all available T
Genome Wide SNP6 array, as of
recent archives (*eg* `broad.mit.`
types was downloaded from the
Each of these segmentation files
During ETL the sample identifer
the SDRF file in the associated m

In order to work with BigQuery,
the name(s) of the table(s) you a

```
import gcp.bigquery a
cn_BQtable = bq.Table
```

From now on, we will refer to thi
table name each time.

Let's start by taking a look at the

```
%bigquery schema --ta
```

name	type
ParticipantBarcode	STRIN
SampleBarcode	STRIN
SampleTypeLetterCode	STRIN
AliquotBarcode	STRIN
Study	STRIN
Platform	STRIN
Chromosome	STRIN
Start	INTE
End	INTE
Num_Probes	INTE
Segment_Mean	FLOA

Unlike most other molecular dat
microRNAs, this data is produce
sizes and positions of these segn

Now we'll use matplotlib to create some simple visual

```
import numpy as np
import matplotlib.pyplot as plt
```

For the segment means, let's invert the log-transform

```
%%sql --module getCNhist

SELECT
  lin_bin,
  COUNT(*) AS n
FROM (
  SELECT
    Segment_Mean,
    (2.*POW(2,Segment_Mean)) AS lin_
    INTEGER(((2.*POW(2,Segment_Mean)
  FROM
    $t
  WHERE
    ( (End-Start+1)>1000 AND SampleT
GROUP BY
  lin_bin
HAVING
  ( n > 2000 )
ORDER BY
  lin_bin ASC
```

```
CNhist = bq.Query(getCNhist,t=cn_BQt
bar_width=0.80
plt.bar(CNhist['lin_bin']+0.1,CNhist
plt.xticks(CNhist['lin_bin']+0.5,CNh
plt.title('Histogram of Average Copy
plt.ylabel('# of segments');
plt.xlabel('integer copy-number');
```

The histogram illustrates that the vast majority of the
either side representing deletions (left) and amplificat

```
  bin
ORDER BY
  bin ASC
```

```
%%sql --module getSLhist_1k_amp

SELECT
  bin,
  COUNT(*) AS n
FROM (
  SELECT
    (END-Start+1) AS segLength,
    INTEGER((END-Start+1)/1000) AS bin
  FROM
    $t
  WHERE
    (END-Start+1)<1000000 AND SampleTypeLetterCode="TP" AND Segment_Mean>0.7 )
GROUP BY
  bin
ORDER BY
  bin ASC
```

```
SLhistDel = bq.Query(getSLhist_1k_del,t=cn_BQtable).results().to_dataframe()
SLhistAmp = bq.Query(getSLhist_1k_amp,t=cn_BQtable).results().to_dataframe()
```

```
plt.plot(SLhist_1k['bin'],SLhist_1k['n'],'ro:');
plt.plot(SLhistDel['bin'],SLhistDel['n'],'bo-')
plt.plot(SLhistAmp['bin'],SLhistDel['n'],'go-',alpha=0.3)
plt.xscale('log');
plt.yscale('log');
plt.xlabel('Segment length (Kb)');
plt.ylabel('# of Segments');
plt.title('Distribution of Segment Lengths');
```

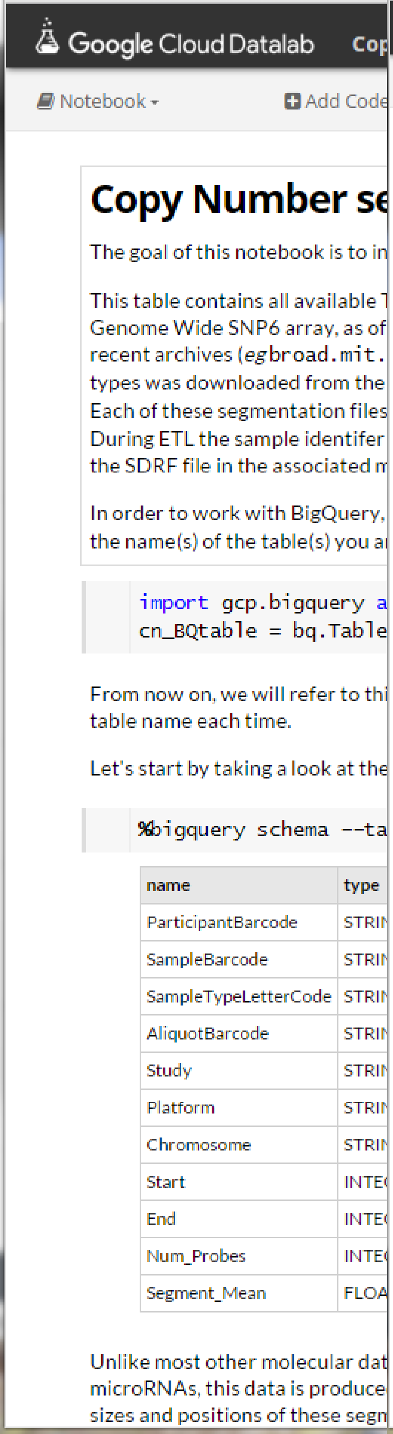
The amplification and deletion distributions are nearly identical and still seem to roughly follow a power-law distribution. We can also infer from this graph that a majority of the segments less than 10Kb in length are either amplifications or deletions, while ~90% of the segments of lengths >100Kb are copy-number neutral.

Help for Python APIs
You can enter `class?` or `member?` within a code cell in the notebook to get help on a Python API.

For example, try `str?` to get help information on the built-in Python method to convert a value to its string representation.

Additional help topics and links are also available from the menu off the Help icon on the top of the page.

Docs and Samples
The Datalab Guide featuring documentation and sample notebooks is also a great way to check out how you can use Datalab.



Copy Number se

The goal of this notebook is to in

This table contains all available T
Genome Wide SNP6 array, as of
recent archives (*eg* `broad.mit.`
types was downloaded from the
Each of these segmentation files
During ETL the sample identifer
the SDRF file in the associated m

In order to work with BigQuery,
the name(s) of the table(s) you a

```
import gcp.bigquery a
cn_BQtable = bq.Table
```

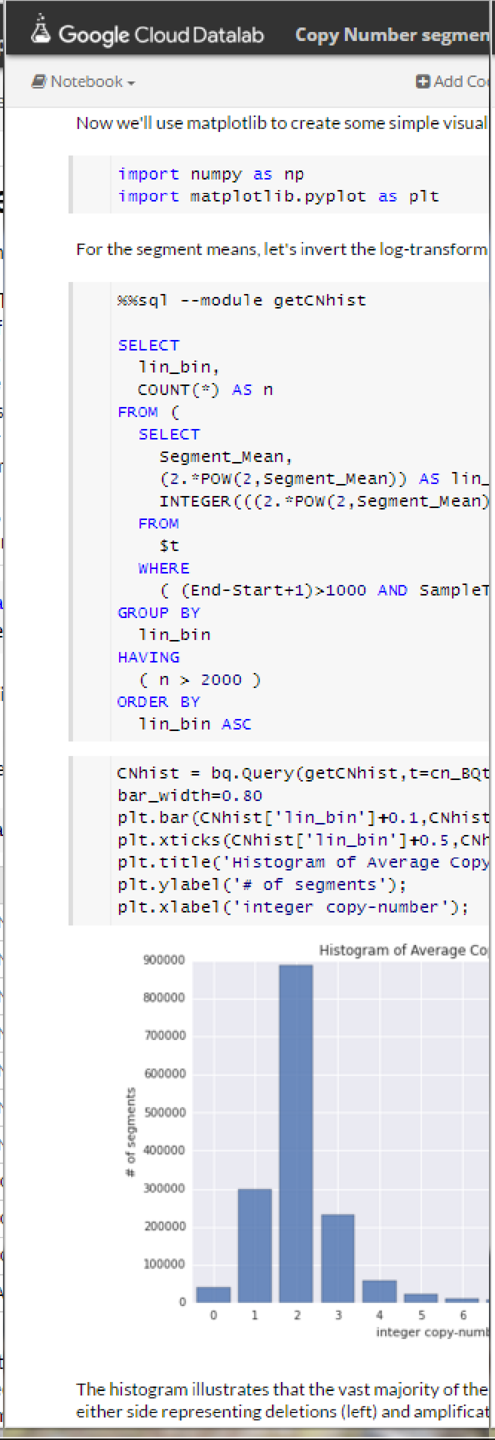
From now on, we will refer to thi
table name each time.

Let's start by taking a look at the

```
%bigquery schema --ta
```

name	type
ParticipantBarcode	STRIN
SampleBarcode	STRIN
SampleTypeLetterCode	STRIN
AliquotBarcode	STRIN
Study	STRIN
Platform	STRIN
Chromosome	STRIN
Start	INTE
End	INTE
Num_Probes	INTE
Segment_Mean	FLOA

Unlike most other molecular dat
microRNAs, this data is produce
sizes and positions of these segn



Now we'll use matplotlib to create some simple visual

```
import numpy as np
import matplotlib.pyplot as plt
```

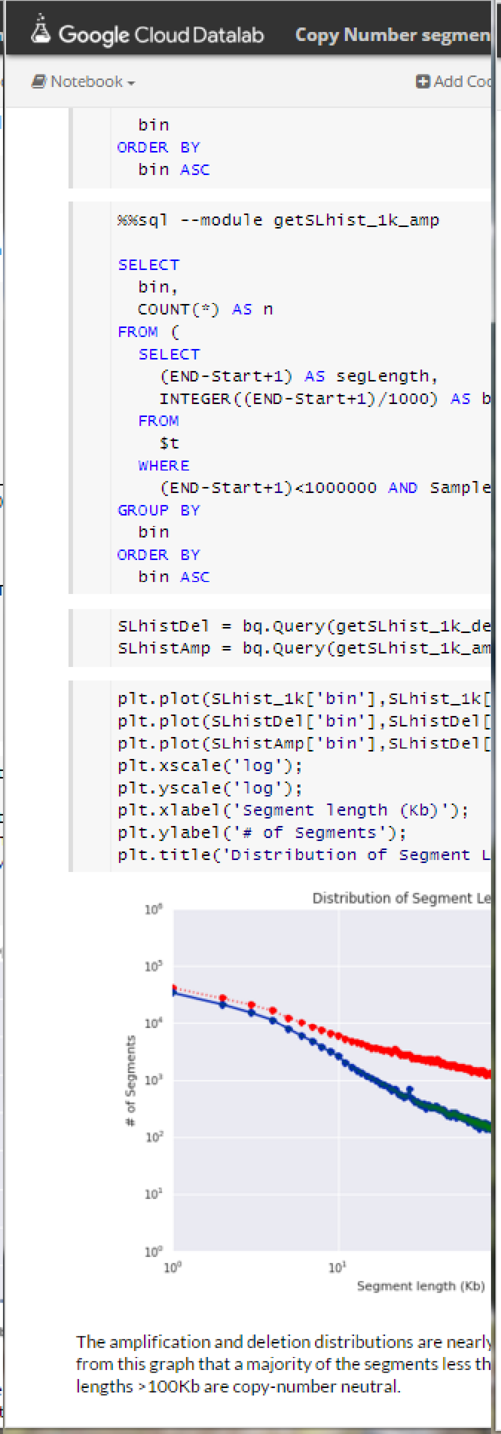
For the segment means, let's invert the log-transform

```
%%sql --module getCNhist

SELECT
  lin_bin,
  COUNT(*) AS n
FROM (
  SELECT
    Segment_Mean,
    (2.*POW(2,Segment_Mean)) AS lin_
    INTEGER(((2.*POW(2,Segment_Mean)
  FROM
    $t
  WHERE
    ( (End-Start+1)>1000 AND SampleT
GROUP BY
  lin_bin
HAVING
  ( n > 2000 )
ORDER BY
  lin_bin ASC
```

```
CNhist = bq.Query(getCNhist,t=cn_BQt
bar_width=0.80
plt.bar(CNhist['lin_bin']+0.1,CNhist
plt.xticks(CNhist['lin_bin']+0.5,CNh
plt.title('Histogram of Average Copy
plt.ylabel('# of segments');
plt.xlabel('integer copy-number');
```

The histogram illustrates that the vast majority of the
either side representing deletions (left) and amplificat



```
  bin
ORDER BY
  bin ASC
```

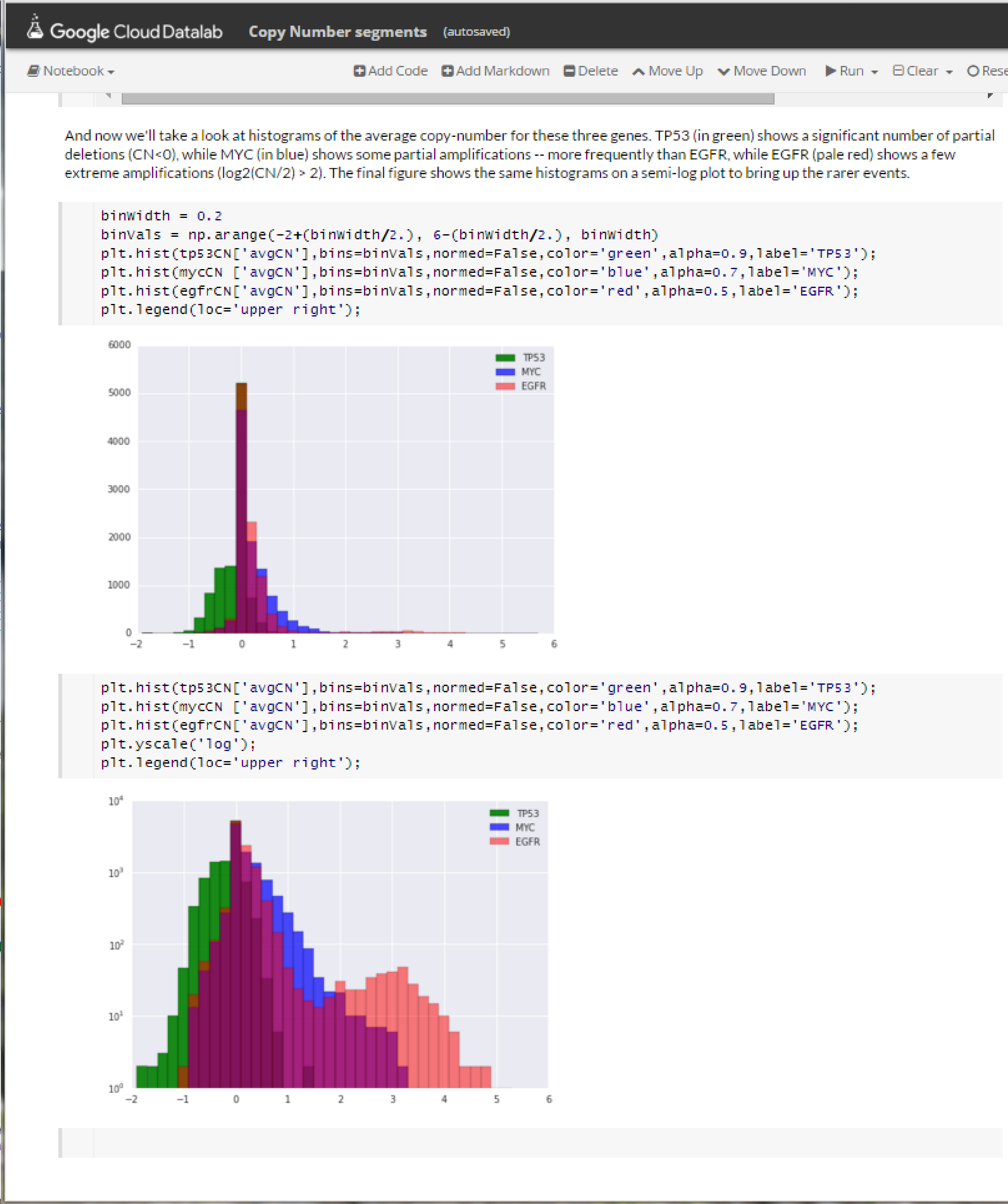
```
%%sql --module getSLhist_1k_amp

SELECT
  bin,
  COUNT(*) AS n
FROM (
  SELECT
    (END-Start+1) AS segLength,
    INTEGER((END-Start+1)/1000) AS b
  FROM
    $t
  WHERE
    (END-Start+1)<1000000 AND Sample
GROUP BY
  bin
ORDER BY
  bin ASC
```

```
SLhistDel = bq.Query(getSLhist_1k_de
SLhistAmp = bq.Query(getSLhist_1k_am
```

```
plt.plot(SLhist_1k['bin'],SLhist_1k[
plt.plot(SLhistDel['bin'],SLhistDel[
plt.plot(SLhistAmp['bin'],SLhistDel[
plt.xscale('log');
plt.yscale('log');
plt.xlabel('Segment length (Kb)');
plt.ylabel('# of Segments');
plt.title('Distribution of Segment L
```

The amplification and deletion distributions are nearly
from this graph that a majority of the segments less th
lengths >100Kb are copy-number neutral.



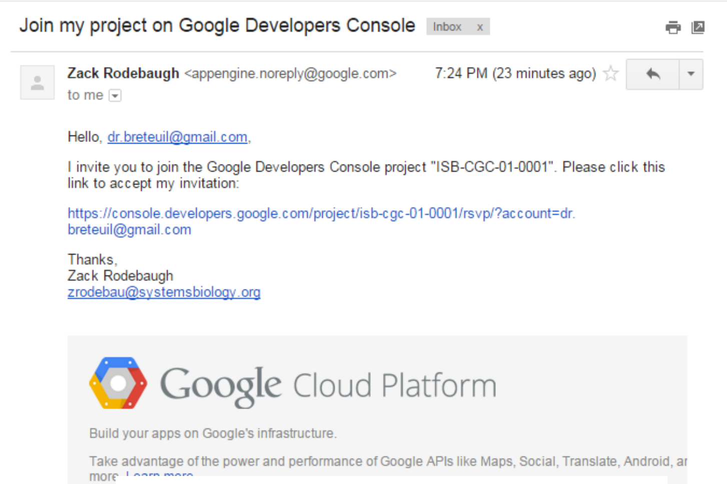
And now we'll take a look at histograms of the average copy-number for these three genes. TP53 (in green) shows a significant number of partial deletions (CN<0), while MYC (in blue) shows some partial amplifications -- more frequently than EGFR, while EGFR (pale red) shows a few extreme amplifications (log2(CN/2) > 2). The final figure shows the same histograms on a semi-log plot to bring up the rarer events.

```
binWidth = 0.2
binVals = np.arange(-2+(binWidth/2.), 6-(binWidth/2.), binWidth)
plt.hist(tp53CN['avgCN'],bins=binVals,normed=False,color='green',alpha=0.9,label='TP53');
plt.hist(mycCN ['avgCN'],bins=binVals,normed=False,color='blue',alpha=0.7,label='MYC');
plt.hist(egfrCN['avgCN'],bins=binVals,normed=False,color='red',alpha=0.5,label='EGFR');
plt.legend(loc='upper right');
```

```
plt.hist(tp53CN['avgCN'],bins=binVals,normed=False,color='green',alpha=0.9,label='TP53');
plt.hist(mycCN ['avgCN'],bins=binVals,normed=False,color='blue',alpha=0.7,label='MYC');
plt.hist(egfrCN['avgCN'],bins=binVals,normed=False,color='red',alpha=0.5,label='EGFR');
plt.yscale('log');
plt.legend(loc='upper right');
```

programmatic access for the Algorithm Developer:

- your own Google Cloud Project , with automatic access to:
  - Cloud Storage
  - BigQuery
  - Google Genomics
  - all Google Compute technologies, including:
    - Compute Engine: anything you can do on your laptop/desktop you can do on a VM
    - Container Engine: fully managed and hosted container orchestration – create and deploy clusters in seconds
    - Dataflow: successor to MapReduce



Join my project on Google Developers Console

Zack Rodebaugh <appengine.noreply@google.com> 7:24 PM (23 minutes ago)
to me

Hello, dr.breteuil@gmail.com,

I invite you to join the Google Developers Console project "ISB-CGC-01-0001". Please click this link to accept my invitation:

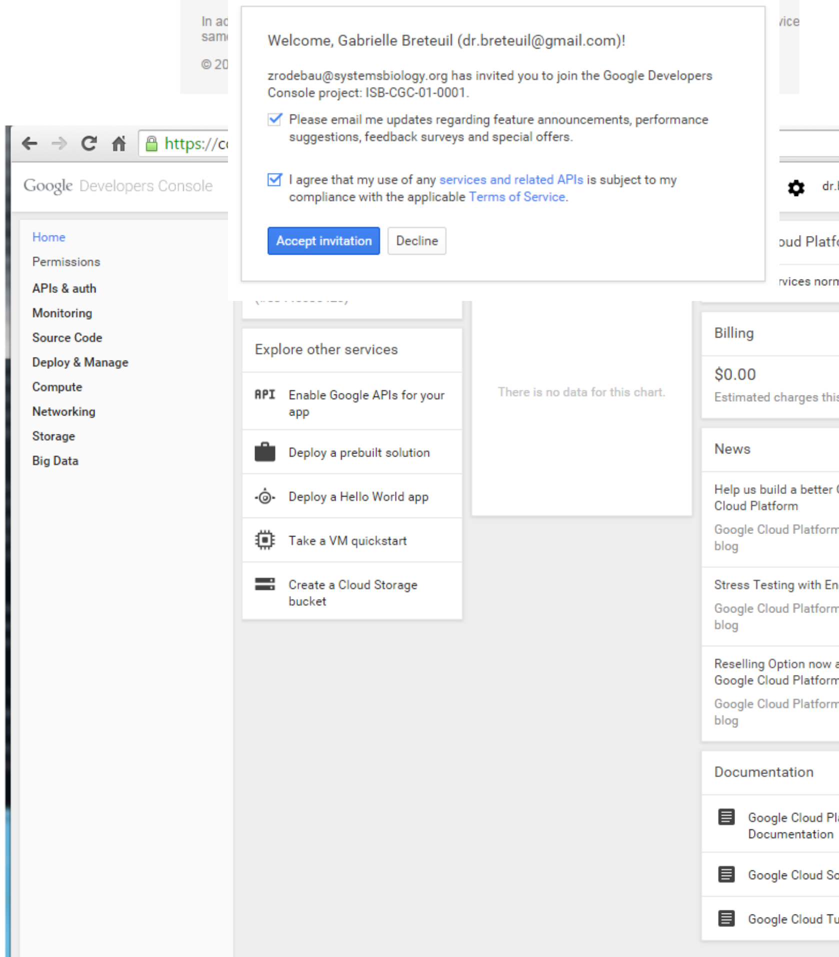
https://console.developers.google.com/project/isb-cgc-01-0001/rsvp/?account=dr.breteuil@gmail.com

Thanks,
Zack Rodebaugh
zrodebau@systemsbiology.org

Google Cloud Platform

Build your apps on Google's infrastructure.

Take advantage of the power and performance of Google APIs like Maps, Social, Translate, Android, ar



Welcome, Gabrielle Breteuil (dr.breteuil@gmail.com)!

zrodebau@systemsbiology.org has invited you to join the Google Developers Console project: ISB-CGC-01-0001.

☑ Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.

☑ I agree that my use of any services and related APIs is subject to my compliance with the applicable Terms of Service.

There is no data for this chart.

Billing

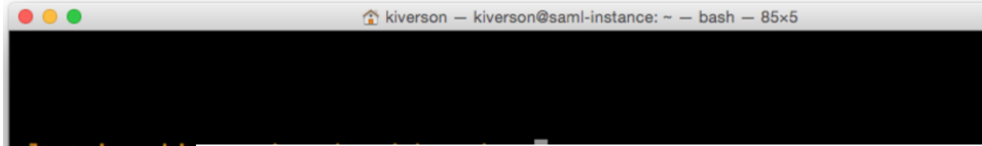
$0.00

Estimated charges this

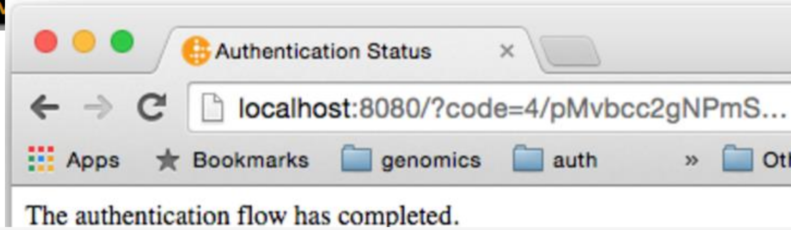
Accept invitation
Decline
https://c
Google Developers Console
Home
Permissions
APIs & auth
Monitoring
Source Code
Deploy & Manage
Compute
Networking
Storage
Big Data
Explore other services
API Enable Google APIs for your app
Deploy a prebuilt solution
Deploy a Hello World app
Take a VM quickstart
Create a Cloud Storage bucket
News
Help us build a better Cloud Platform
Google Cloud Platform blog
Stress Testing with En
Google Cloud Platform blog
Reselling Option now a Google Cloud Platform
Google Cloud Platform blog
Documentation
Google Cloud Pl Documentation
Google Cloud So
Google Cloud Tu

the ISB-CGC API provides programmatic access to the same functionality as the web-app and more:

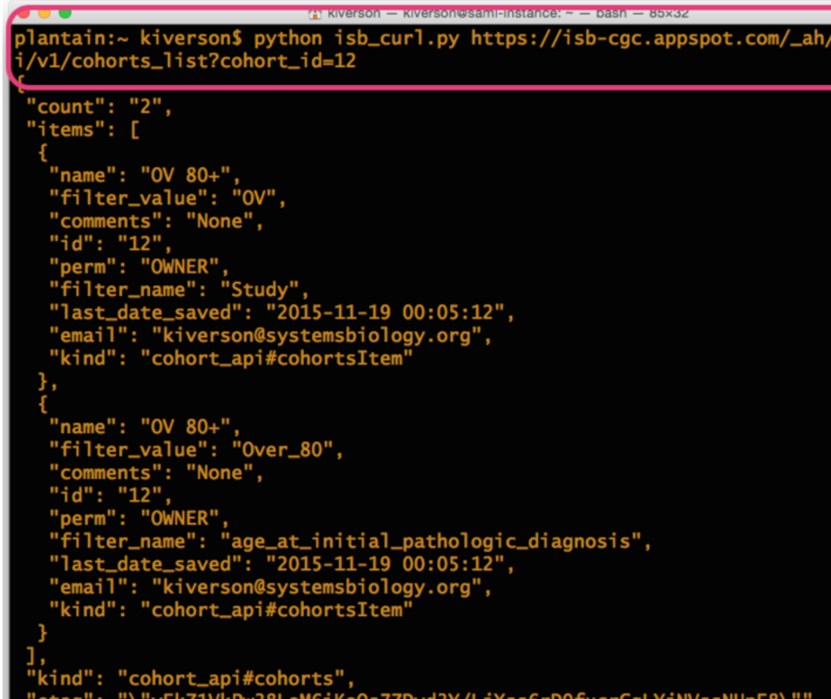
- Cloud Endpoints API (backed by App Engine)
    - authenticate from the command-line
    - make requests to Endpoints API, *eg*:
        - get list of my cohorts
        - get cohort details
        - save a new cohort
        - get list of data files associated with a cohort



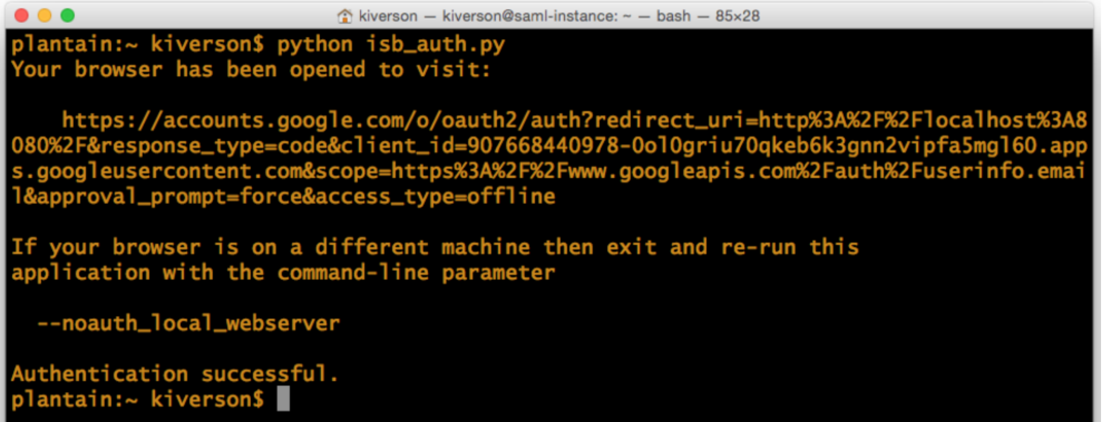
```
plantain:~ ki
```



The authentication flow has completed.



```
plantain:~ kiverson$ python isb_curl.py https://isb-cgc.appspot.com/_ah/
i/v1/cohorts_list?cohort_id=12
{
 "count": "2",
 "items": [
  {
   "name": "OV 80+",
   "filter_value": "OV",
   "comments": "None",
   "id": "12",
   "perm": "OWNER",
   "filter_name": "Study",
   "last_date_saved": "2015-11-19 00:05:12",
   "email": "kiverson@systemsbiology.org",
   "kind": "cohort_api#cohortsItem"
  },
  {
   "name": "OV 80+",
   "filter_value": "Over_80",
   "comments": "None",
   "id": "12",
   "perm": "OWNER",
   "filter_name": "age_at_initial_pathologic_diagnosis",
   "last_date_saved": "2015-11-19 00:05:12",
   "email": "kiverson@systemsbiology.org",
   "kind": "cohort_api#cohortsItem"
  }
 ],
 "kind": "cohort_api#cohorts",
```



```
plantain:~ kiverson$ python isb_auth.py
Your browser has been opened to visit:

    https://accounts.google.com/o/oauth2/auth?redirect_uri=http%3A%2F%2Flocalhost%3A8
080%2F&response_type=code&client_id=907668440978-0ol0griu70qkeb6k3gnn2vipfa5mg160.app
s.googleusercontent.com&scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.emai
l&approval_prompt=force&access_type=offline

If your browser is on a different machine then exit and re-run this
application with the command-line parameter

  --noauth_local_webserver

Authentication successful.
plantain:~ kiverson$ █
```

Summary

ISB-CGC Phase 1

- Low-level sequence and SNP array data as ***files*** in **Cloud Storage**
- High-level data and annotations as ***tables*** in **BigQuery**
- Multiple access modes and interfaces:
  - Interactive web-application
  - R, Python, SQL, and JavaScript
  - Endpoint APIs

ISB-CGC Phase 2

- Low-level sequence data in **Google Genomics**
- Variant calls in **Google Genomics** and **BigQuery**

NATIONAL CANCER INSTITUTE®

TCGA

CBIIT Center for Biomedical Informatics and Information Technology

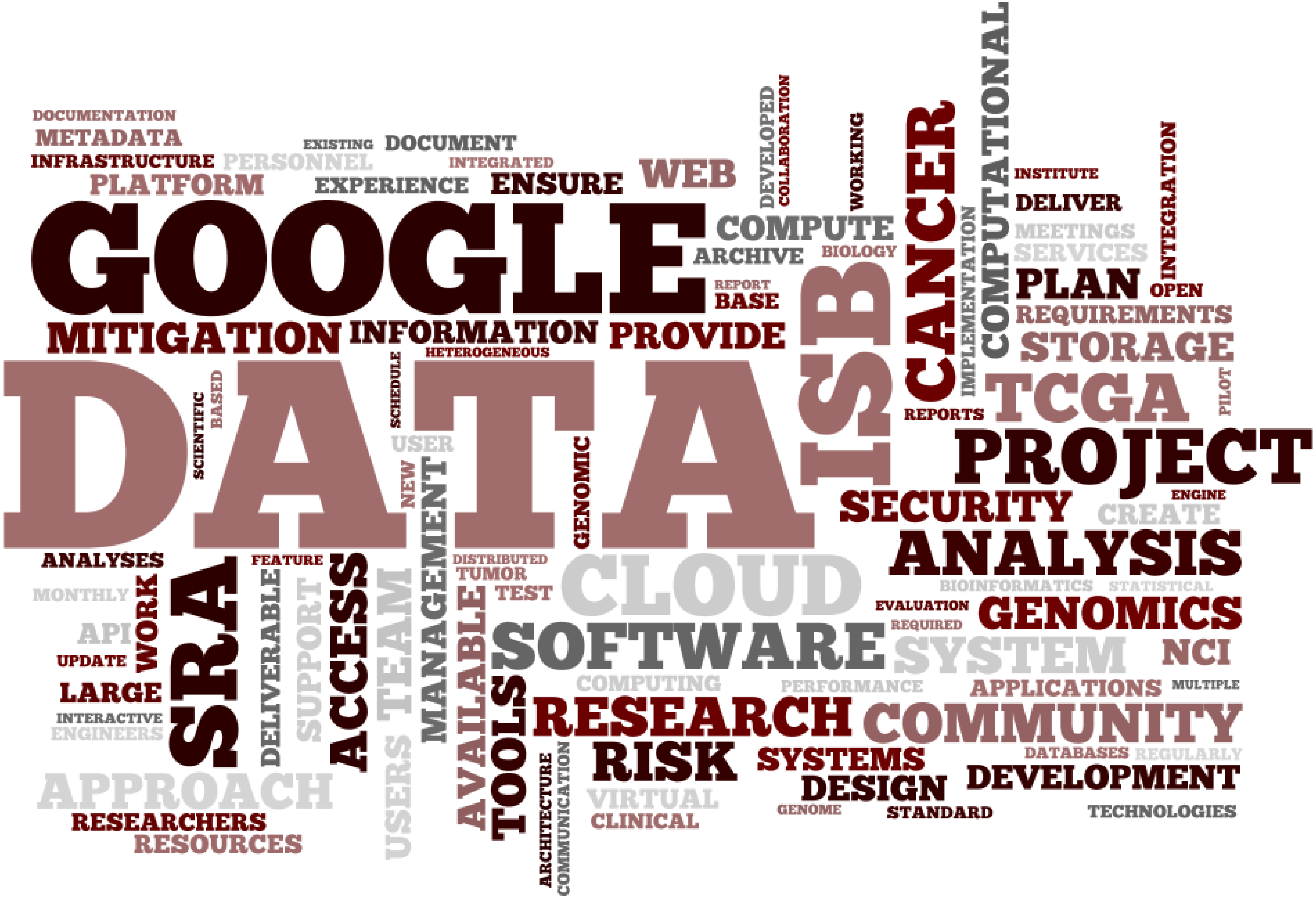
NCI | NHGRI

This project has been funded in whole with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400007C.

ISB Cancer Genomics Cloud

Questions?

www.isb-cgc.org
info@isb-cgc.org


DOCUMENTATION
METADATA
INFRASTRUCTURE
PERSONNEL
PLATFORM
EXISTING
DOCUMENT
INTEGRATED
EXPERIENCE
ENSURE
WEB
DEVELOPED
COLLABORATION
WORKING
CANCER
COMPUTATIONAL
INSTITUTE
DELIVER
INTEGRATION
GOOGLE
COMPUTE
ARCHIVE
BIOLOGY
IMPLEMENTATION
MEETINGS
SERVICES
REPORT
BASE
PLAN
OPEN
REQUIREMENTS
MITIGATION
INFORMATION
PROVIDE
ISB
STORAGE
HETEROGENEOUS
DATA
SCHEDULE
SCIENTIFIC
BASED
TCGA
PILOT
REPORTS
USER
NEW
MANAGEMENT
GENOMIC
PROJECT
ENGINE
SECURITY
CREATE
ANALYSIS
ANALYSES
FEATURE
DISTRIBUTED
TUMOR
TEST
CLOUD
BIOINFORMATICS
STATISTICAL
MONTHLY
WORK
SRA
DELIVERABLE
SUPPORT
ACCESS
TEAM
EVALUATION
GENOMICS
REQUIRED
API
SOFTWARE
SYSTEM
NCI
UPDATE
AVAILABLE
LARGE
COMPUTING
PERFORMANCE
APPLICATIONS
MULTIPLE
INTERACTIVE
ENGINEERS
TOOLS
RESEARCH
COMMUNITY
USERS
RISK
SYSTEMS
DATABASES
REGULARLY
APPROACH
ARCHITECTURE
COMMUNICATION
VIRTUAL
DESIGN
DEVELOPMENT
RESEARCHERS
CLINICAL
GENOME
STANDARD
TECHNOLOGIES
RESOURCES