The ISB Cancer Genomics Cloud

Sheila M Reynolds
Institute for Systems Biology
April 6th 2016

Bio IT World Conference & Expo

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Google Genomics

CSRA

ISB-CGC Team Members

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ilya Shmulevich
Sheila Reynolds
Phyliss Lee
Michael Miller
Kelly Iverson
Abigail Hahn
Zack Rodebaugh
Kalle Leinonen
Dave Gibbs
Varsha Dhankani

Google Genomics

Jonathan Bingham
Nicole Deflaux
Matt Bookman
Jaclyn Kollar

CSRA

David Pot
Ross Casanova
Sandeep Namburi
Madelyn Reyes

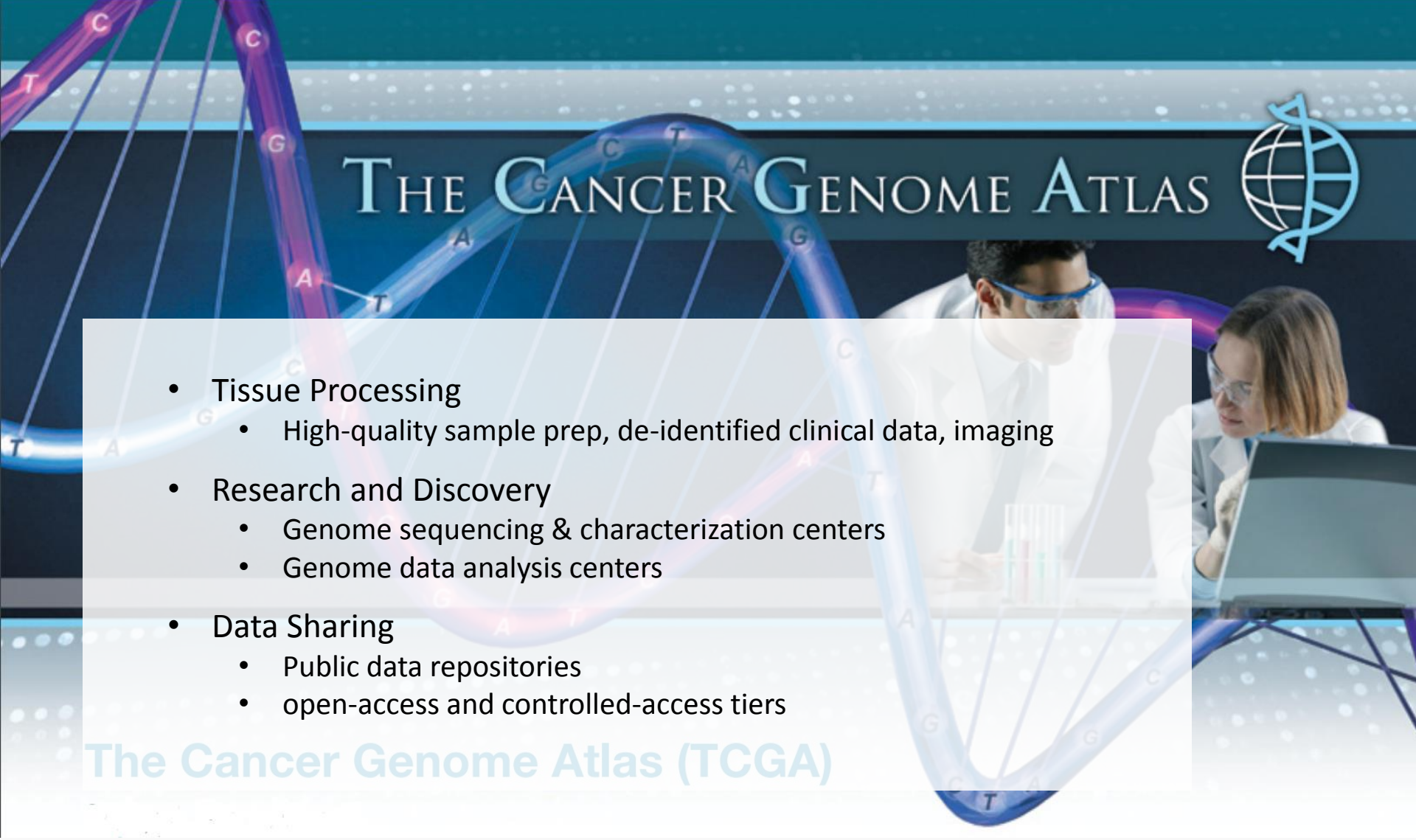
This project has been funded in whole with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400007C.

The Cancer Genome Atlas (TCGA)

" … a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing"

http://cancergenome.nih.gov/

The Cancer Genome Atlas



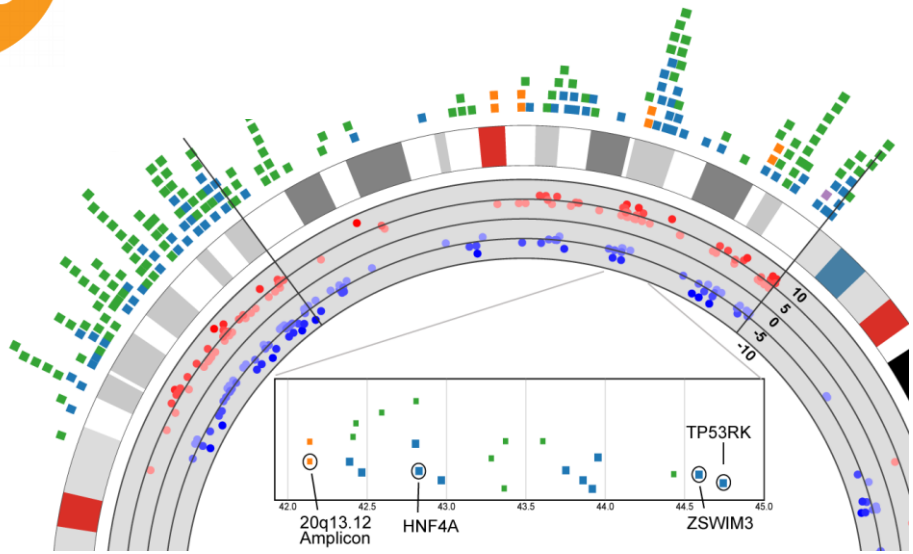
- Tissue Processing
  - High-quality sample prep, de-identified clinical data, imaging
- Research and Discovery
  - Genome sequencing & characterization centers
  - Genome data analysis centers
- Data Sharing
  - Public data repositories
  - open-access and controlled-access tiers

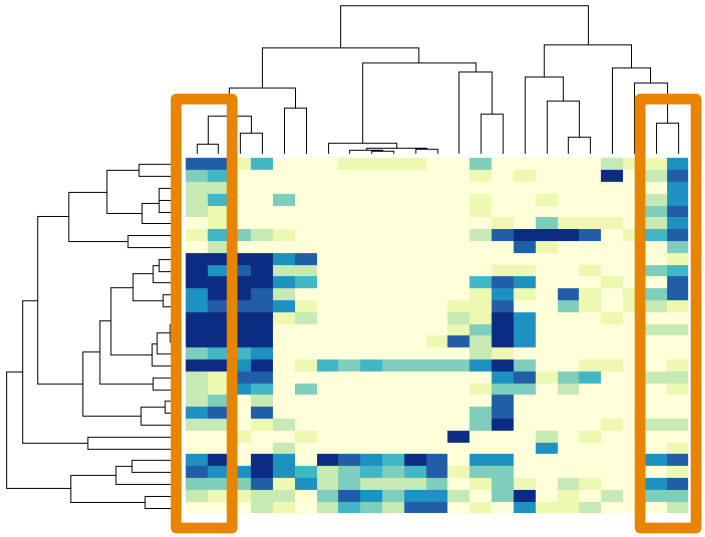
The Cancer Genome Atlas (TCGA)

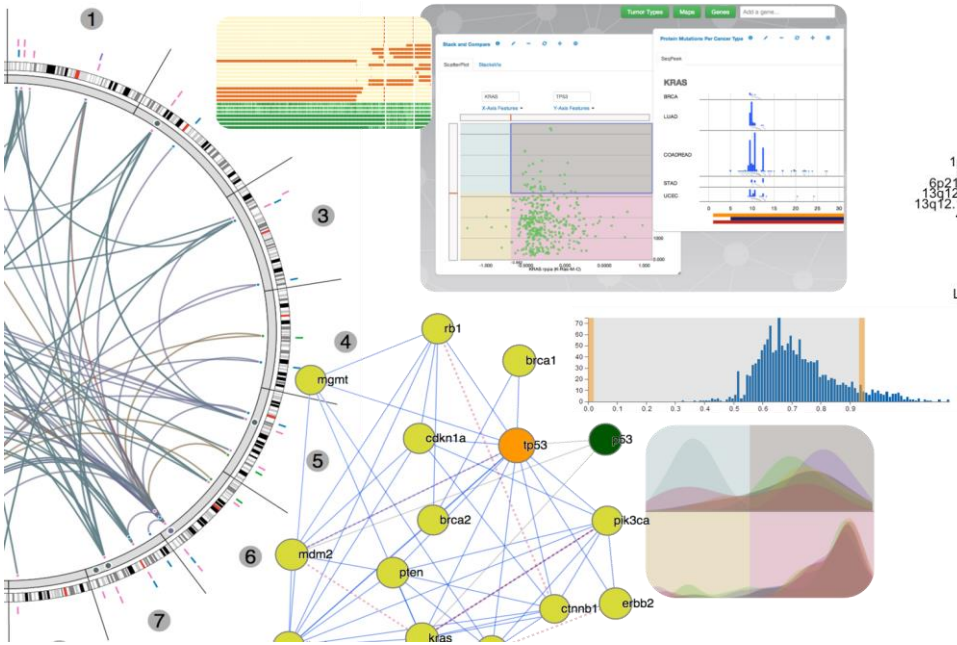
" … a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing"

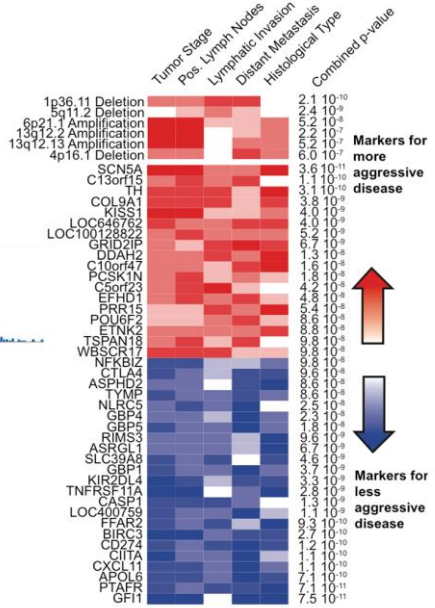
http://cancergenome.nih.gov/

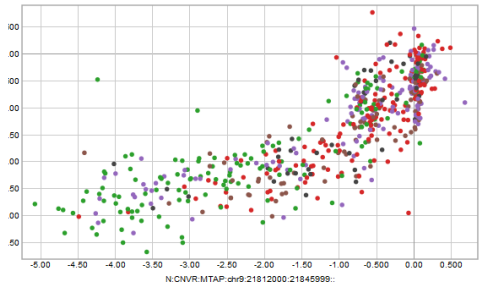
ISB Genome Data Analysis Center

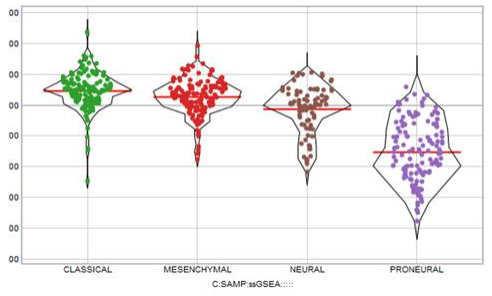




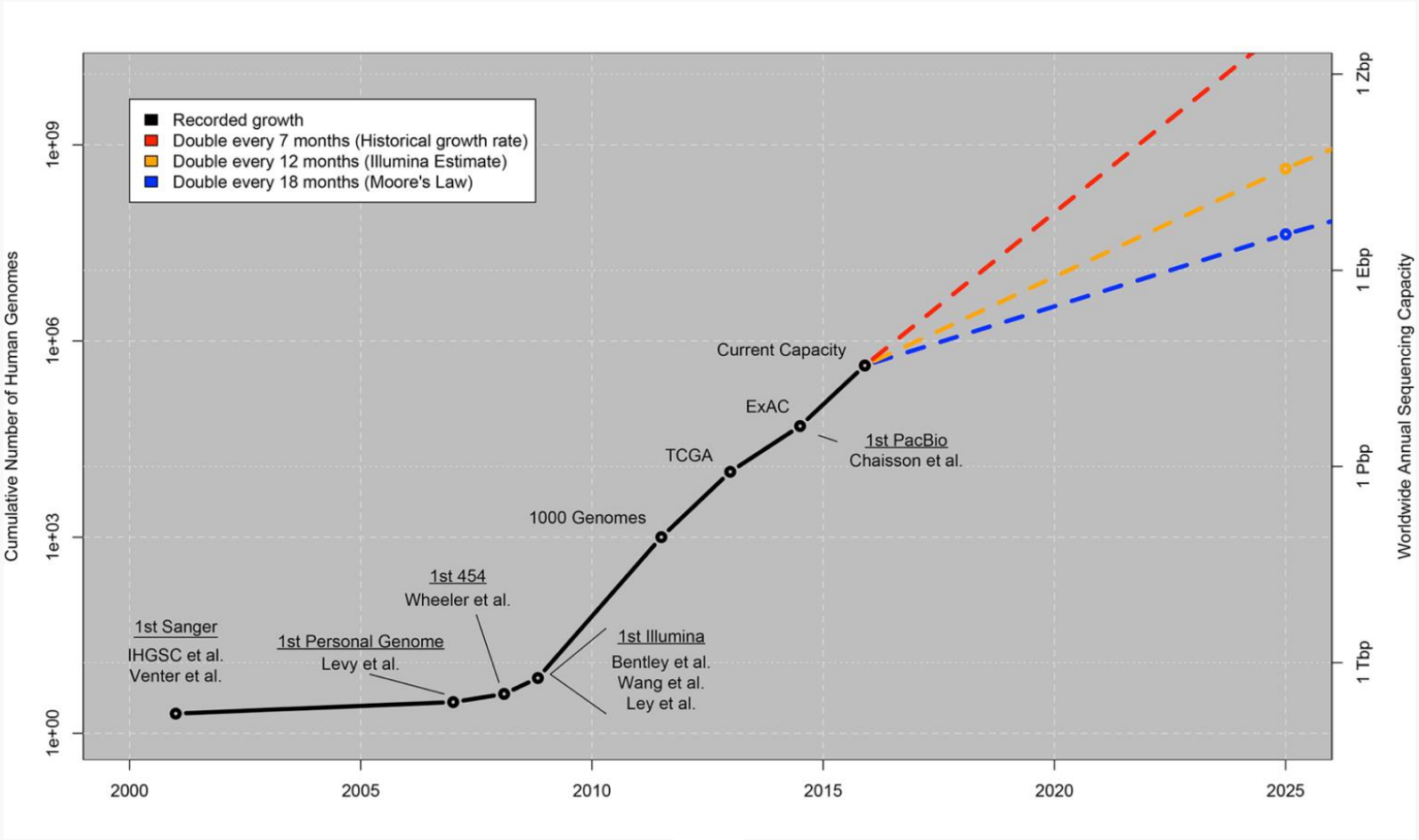






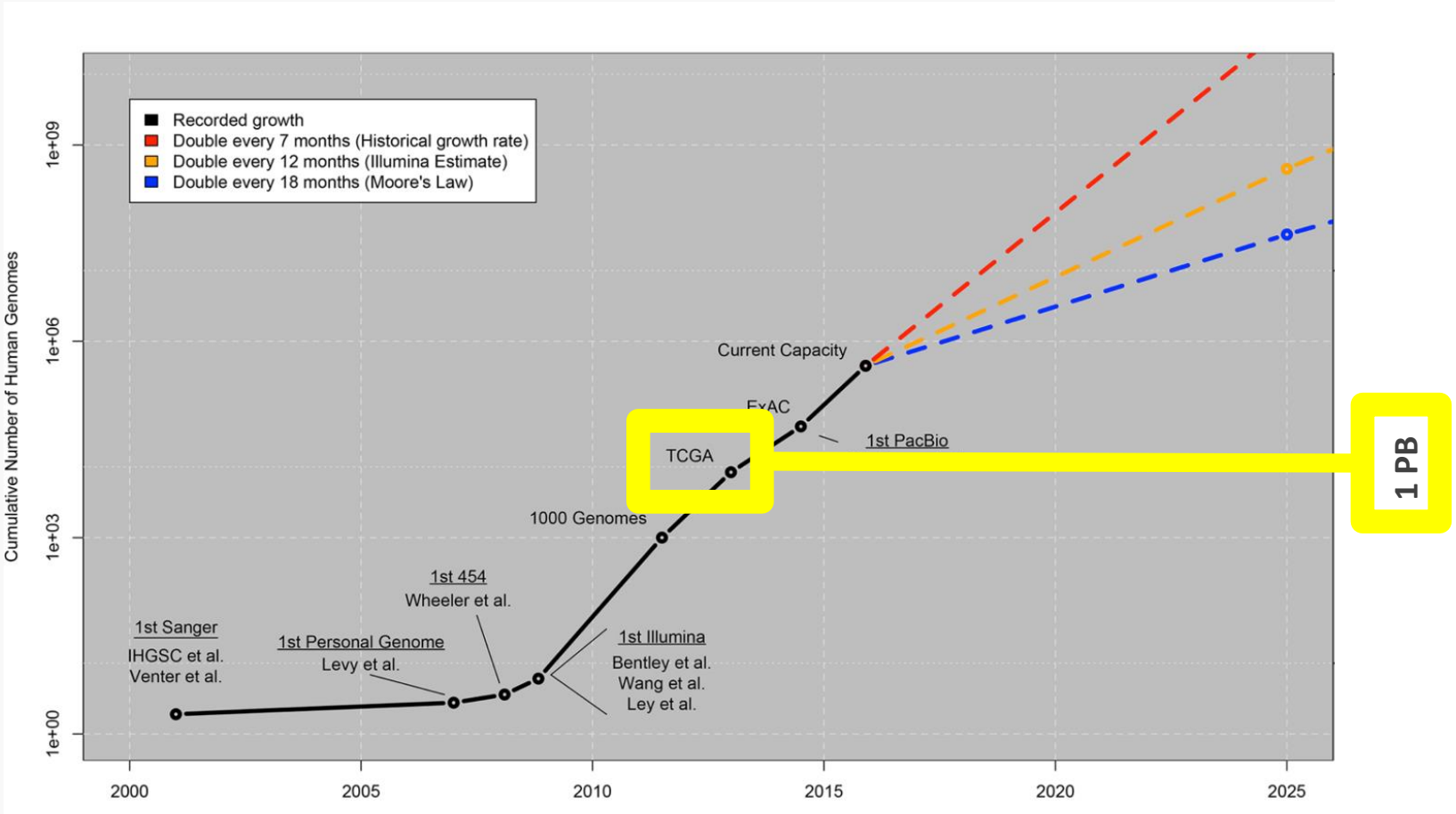


The Challenge of Big Data



Big Data: Astronomical or Genomical? Stephens *et al*, PLOS Biology, July 2015

The Challenge of Big Data



Big Data: Astronomical or Genomical? Stephens *et al*, PLOS Biology, July 2015

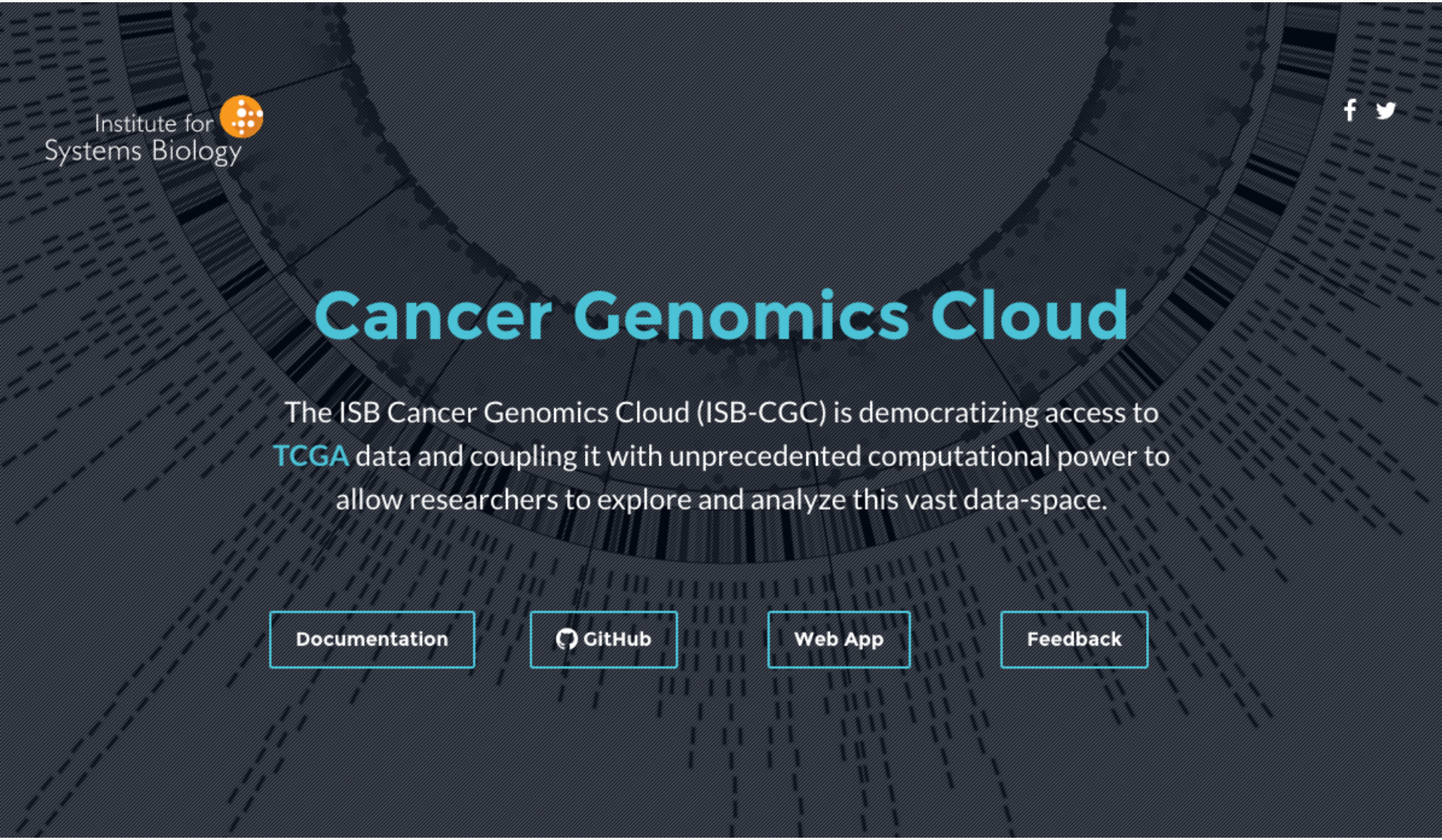
NCI Cancer Genomics Cloud Pilots

Bringing data and computation together to create knowledge
that accelerates cancer research and enables precision medicine

- Goals:
  - Explore innovative methods for accessing and computing on genomic data
  - Democratize access to NCI-generated data sets
  - Cost-effective computational support to the cancer research community
- Timeline:
  - 2015: Design and Build
  - 2016: Community Evaluation

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Source: https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots



www.isb-cgc.org

Cloud Paradigm Shift(s)

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

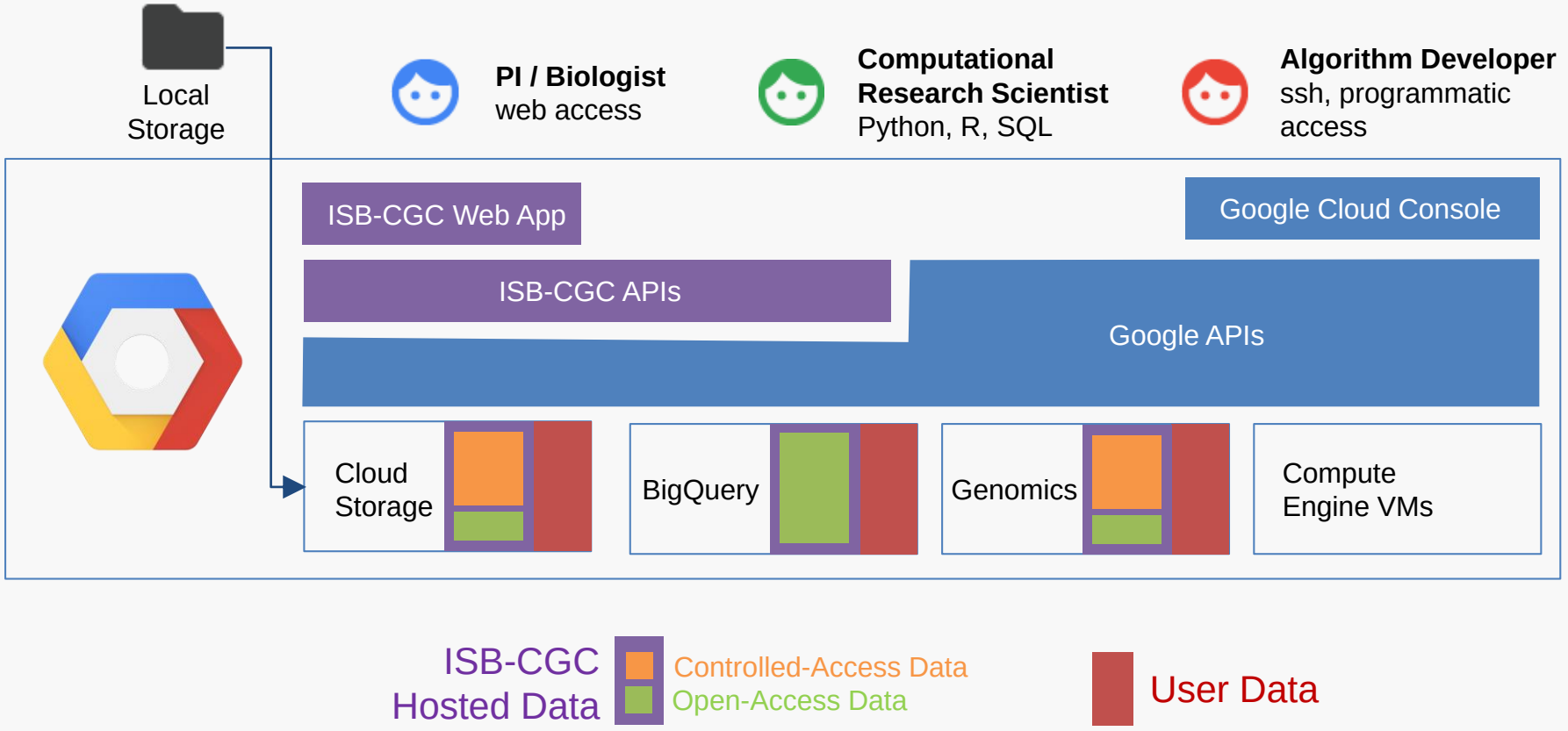
- Cloud-aware computing
  - rethink/redevelop approaches to fully leverage the power of the cloud
  - massively parallel, bursty, opportunistic computing

Our Mission

… is 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

Our Approach: build an open platform that can grow and evolve to satisfy a broad range of users and use-cases



Data as a Service

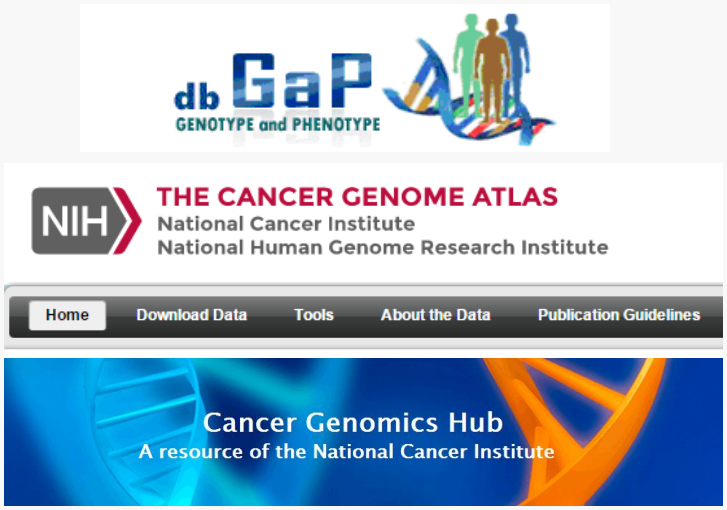
Phase 1

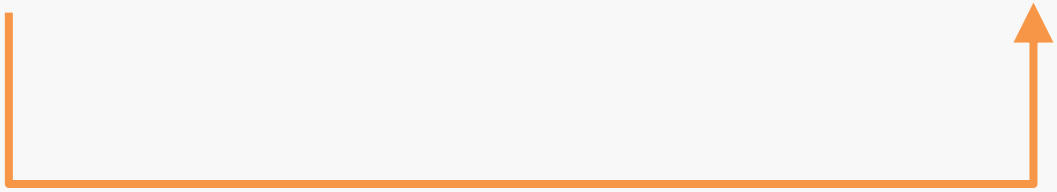
- Low-level sequence and SNP data as ***files*** in **Cloud Storage**
- High-level data and annotations as ***tables*** in **BigQuery**

Phase 2

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

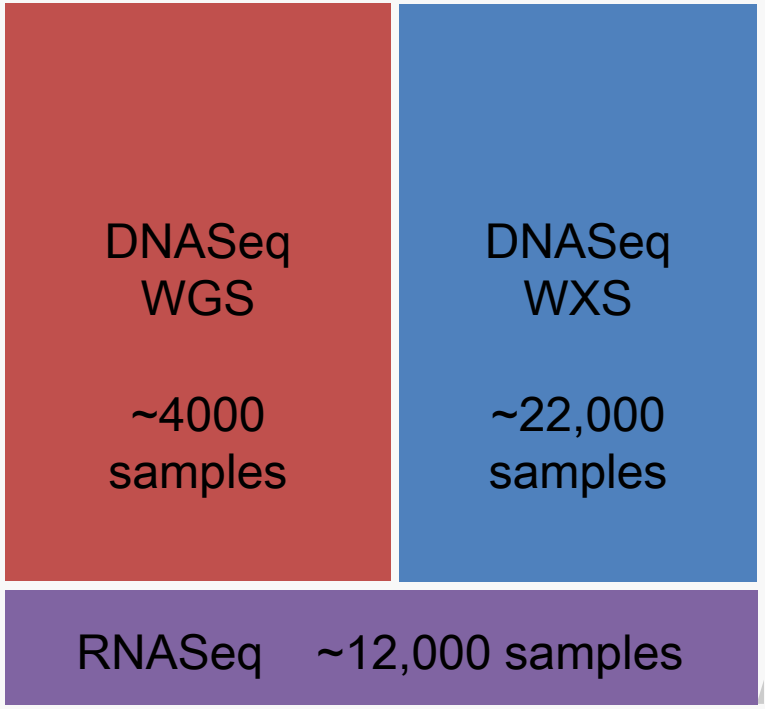
TCGA Data in the Cloud





TCGA Size

>1 PB of sequence data
(controlled access)



TCGA Size & Complexity

>1 PB of sequence data
(controlled access)

~400,000 files of
heterogeneous data
(mostly open-access)

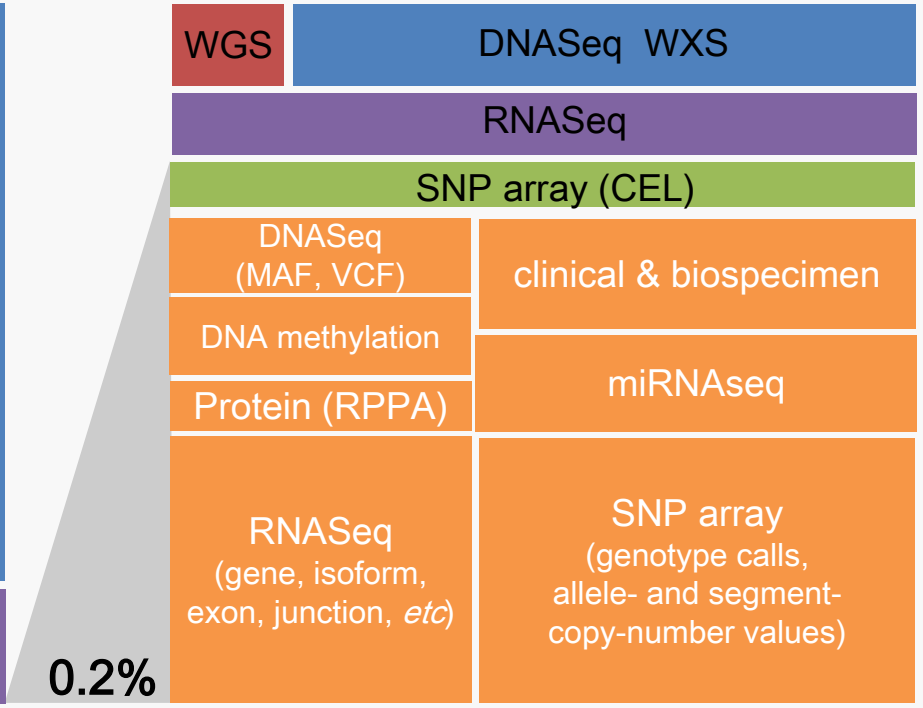


Table Details: Clinical_data

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_neoplasi			
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_follo			
days_to_initial_pat			
days_to_submitted			
ethnicity			
frozen_specimen_i			
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_Componen			
GO_Molecular_Function			
Gene_Type			
Genome_Change			

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			
max_percei			
max_percei			
max_percei			
max_percei			
max_percei			
max_percei			
max_percei			

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...

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+b
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 be

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	NU	
Study	STRING	NU	
itemTypeName	STRING	NU	
itemBarcode	STRING	NU	
AliquotBarcode	STRING	NU	
ParticipantBarcode	STRING	NU	
SampleBarcode	STRING	NU	
dateAdded	STRING	NU	
dateCreated	STRING	NU	
dateEdited	STRING	NU	

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...
	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...

Table Details: mRNA_UNC_HiSeq_RSEM

Schema

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

Table Details: mRNA_BCGSC_HiSeq_RPKM

Schema

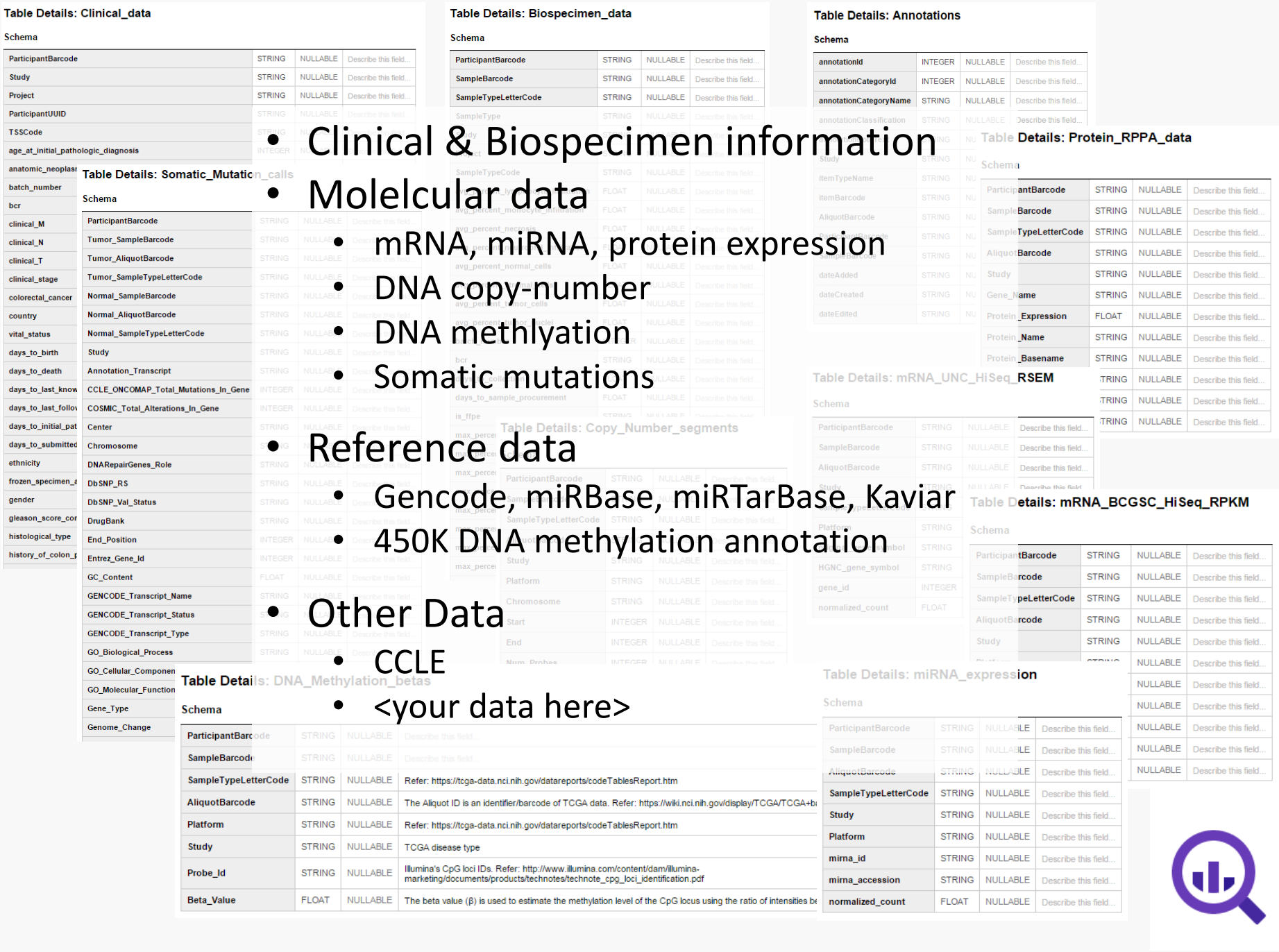
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
		NULLABLE	Describe this field...
		NULLABLE	Describe this field...
		NULLABLE	Describe this field...
		NULLABLE	Describe this field...
		NULLABLE	Describe this field...
		NULLABLE	Describe this field...

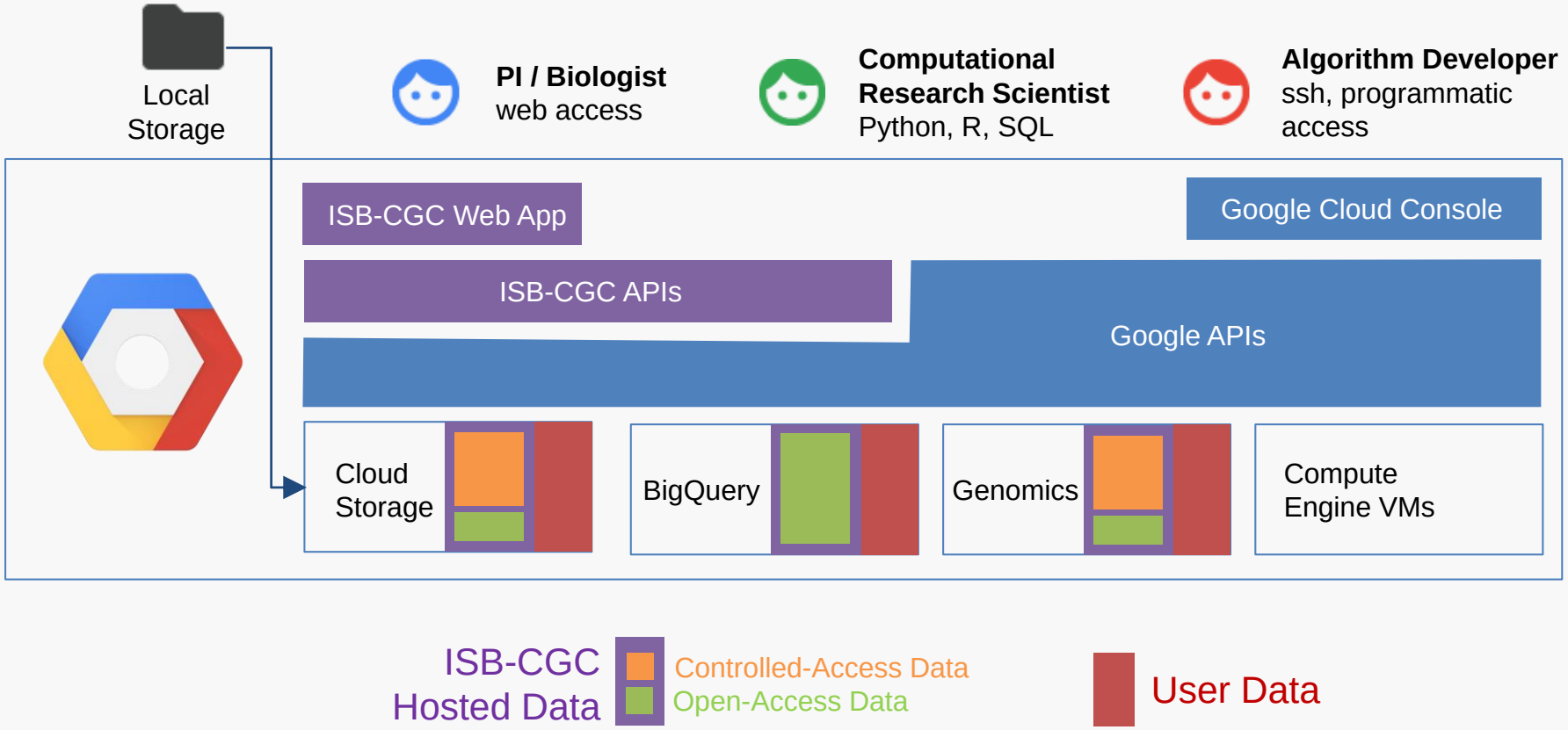
Table Details: miRNA_expression

Schema

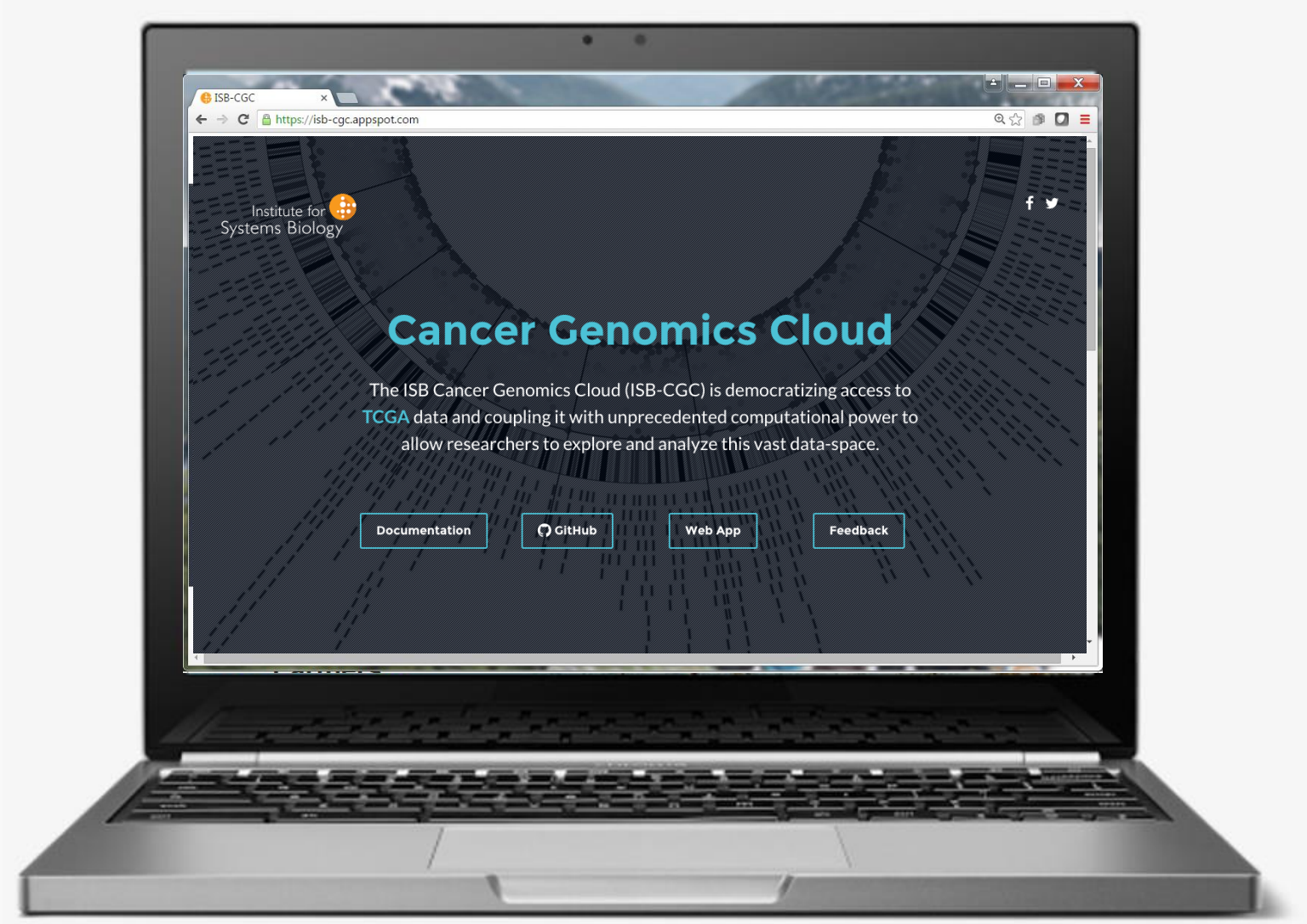
ParticipantBarcode	STRING	NULLABLE	Describe this field...
SampleBarcode	STRING	NULLABLE	Describe this field...
AliquotBarcode	STRING	NULLABLE	Describe this field...
SampleTypeLetterCode	STRING	NULLABLE	Describe this field...
Study	STRING	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe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

- Clinical & Biospecimen information
- Molelcular data
  - mRNA, miRNA, protein expression
  - DNA copy-number
  - DNA methlyation
  - Somatic mutations
- Reference data
  - Gencode, miRBase, miRTarBase, Kaviar
  - 450K DNA methylation annotation
- Other Data
  - CCLE
  - <your data here>

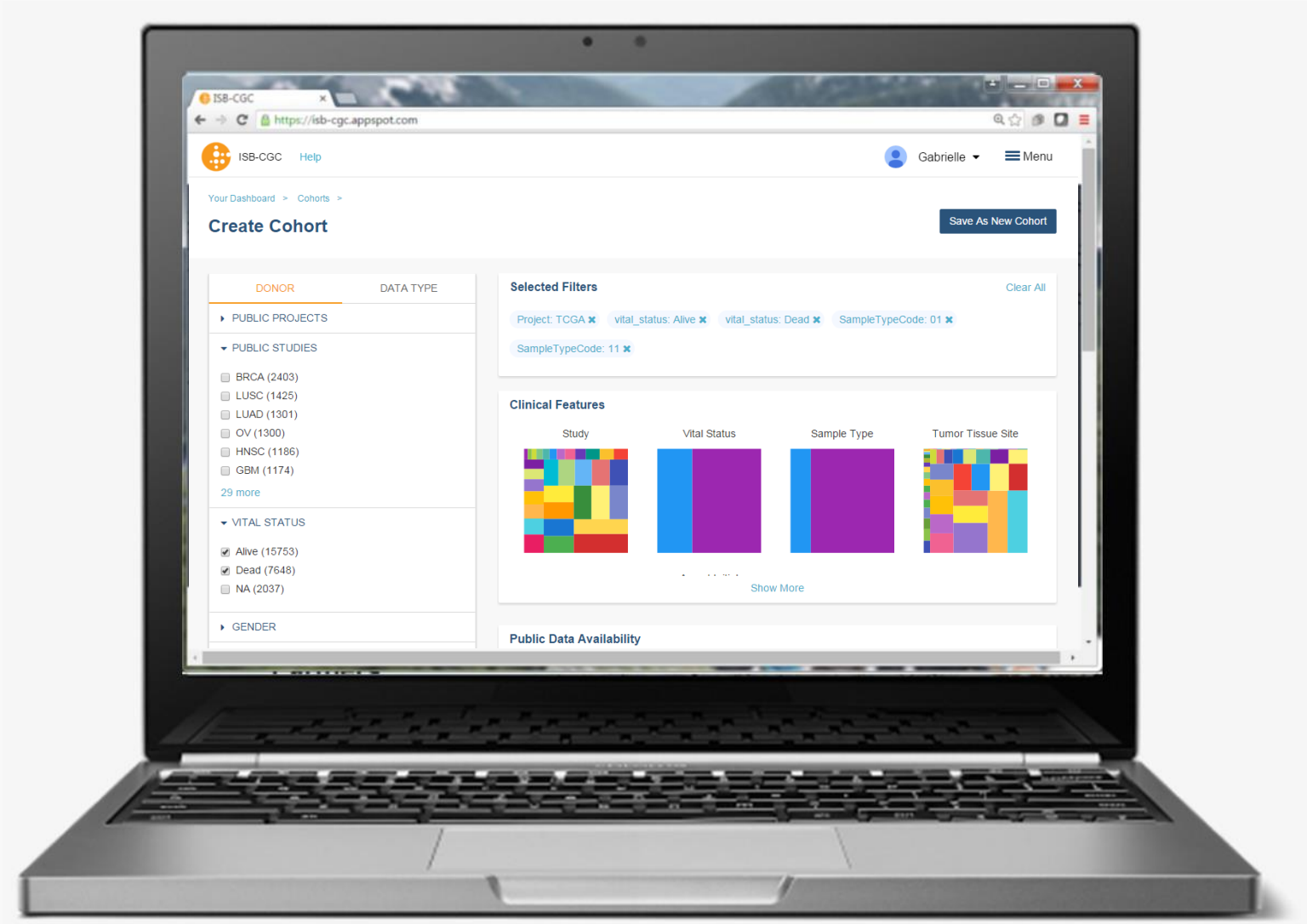



Local Storage
PI / Biologist
web access
Computational Research Scientist
Python, R, SQL
Algorithm Developer
ssh, programmatic access
ISB-CGC Web App
Google Cloud Console
ISB-CGC APIs
Google APIs
Cloud Storage
BigQuery
Genomics
Compute Engine VMs
ISB-CGC Hosted Data
Controlled-Access Data
Open-Access Data
User Data

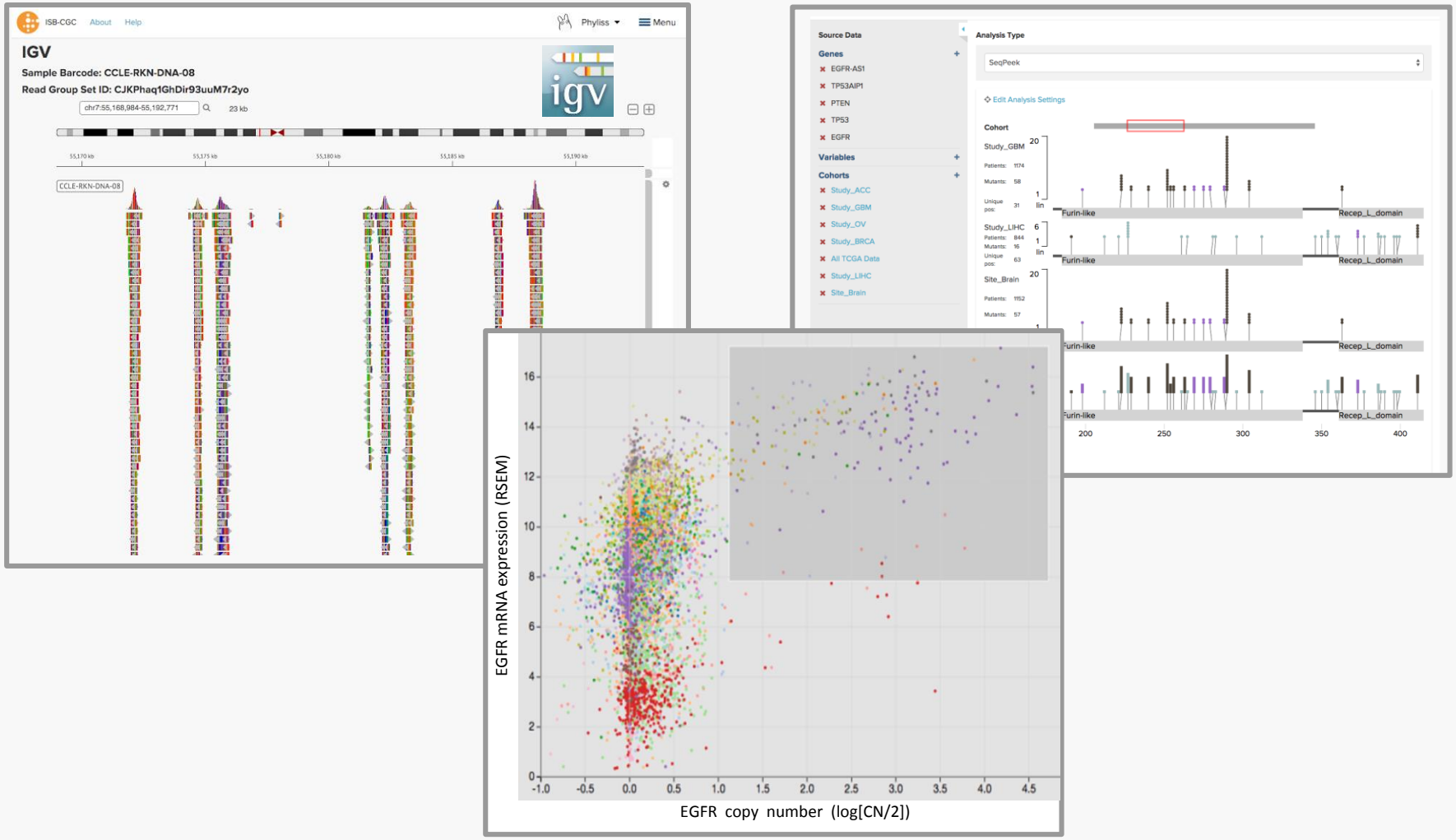
Web access for the PI / Biologist:



custom, interactive cohort definition



Interactive data visualizations:



Python, R, and SQL for the Computational Scientist:

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

Repositories | People 32 | Teams 5 | Settings

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

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

Repositories | People 32 | Teams 5 | Settings

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-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 genomic data, correlating 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 examples of using builtin BigQuery functions like Pearson correlation, or even how to implement more complex functions like Spearmans correlation. Queries can be simple character vectors, or standalone files. Results are returned as data frames using the 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 dispatched with arguments using the DisplayAndDispatchQuery function, found in the file of the same name in examples-R/R.

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

The Comprehensive R Archive Network

R Client for Google Genomics API

Bioconductor version: Release (3.2)

Provides an R package to interact with the Google Genomics API.

Author: Cassie Doll [aut], Nicole Deflaux [aut], Siddhartha Bagaria [aut, cre]

bigrquery: An Interface to Google's BigQuery API

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), dplyr (≥ 0.3.0)
Suggests:	testthat
Published:	2015-01-13
Author:	Hadley Wickham [aut, cre], RStudio [cph]

RStudio®

SQL

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

Repositories | People 32 | Teams 5 | Settings

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

IP[y]: IPython
Interactive Computing

CLOUD DATALAB BETA

An easy to use interactive tool for large-scale data exploration, analysis, and visualization.

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.

SQL

Google Cloud Datalab

Notebook ▾ Add Code

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 (*egbroad.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	STRI
SampleBarcode	STRI
SampleTypeLetterCode	STRI
AliquotBarcode	STRI
Study	STRI
Platform	STRI
Chromosome	STRI
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

Google Cloud Datalab Copy Number segmen

Notebook ▾ Add Co

Now we'll use matplotlib to create some simple visual

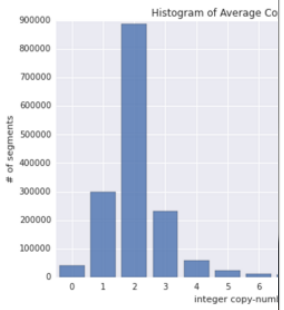
```
import numpy as np
import matplotlib.pyplot as plt
```

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

Google Cloud Datalab Copy Number segmen

Notebook ▾ Add Co

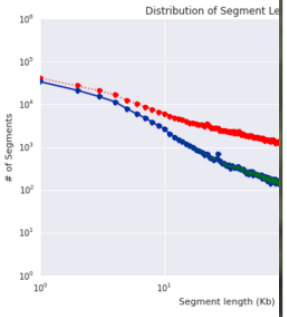
```
  bin
ORDER BY
  bin ASC
```

```
%%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
```



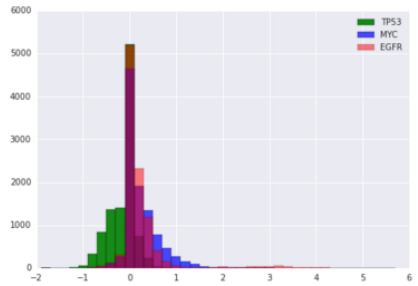
The amplification and deletion distributions are nearly
from this graph that a majority of the segments less th
lengths >100Kb are copy-number neutral.

Google Cloud Datalab Copy Number segments (autosaved)

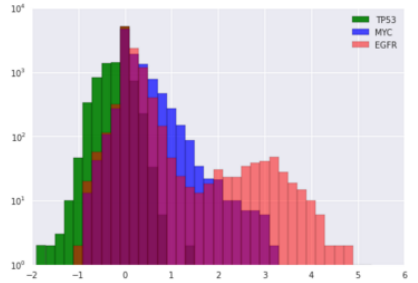
Notebook ▾ Add Code Add Markdown Delete Move Up Move Down Run ▾ Clear ▾ Rese

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');
```



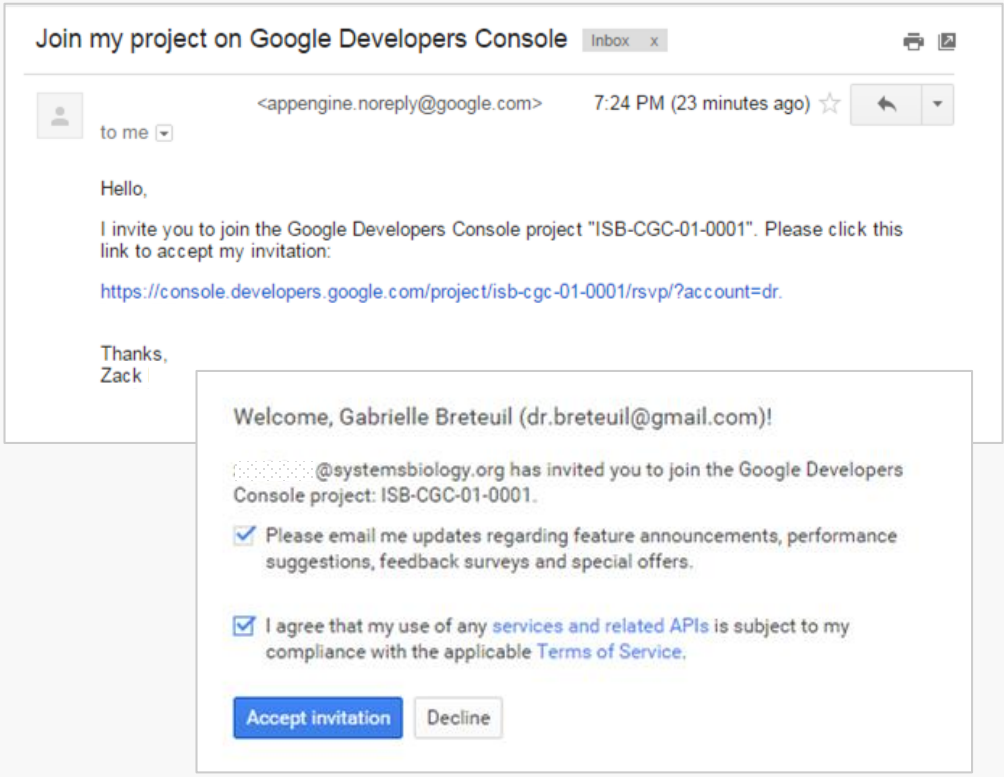

```
plt.hist(tp53CN['avgCN'],bins=binVals,normed=False,color='green',alpha=0.9,label='TP53');
plt.hist(mycCN ['avgCN'],bins=binVals,normed=False,color='blue',alpha=0.7,label='MYC');
plt.hist(egfrCN['avgCN'],bins=binVals,normed=False,color='red',alpha=0.5,label='EGFR');
plt.yscale('log');
plt.legend(loc='upper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



Programmatic access for the Algorithm Developer:

Your own Google Cloud Project, with automatic access to:

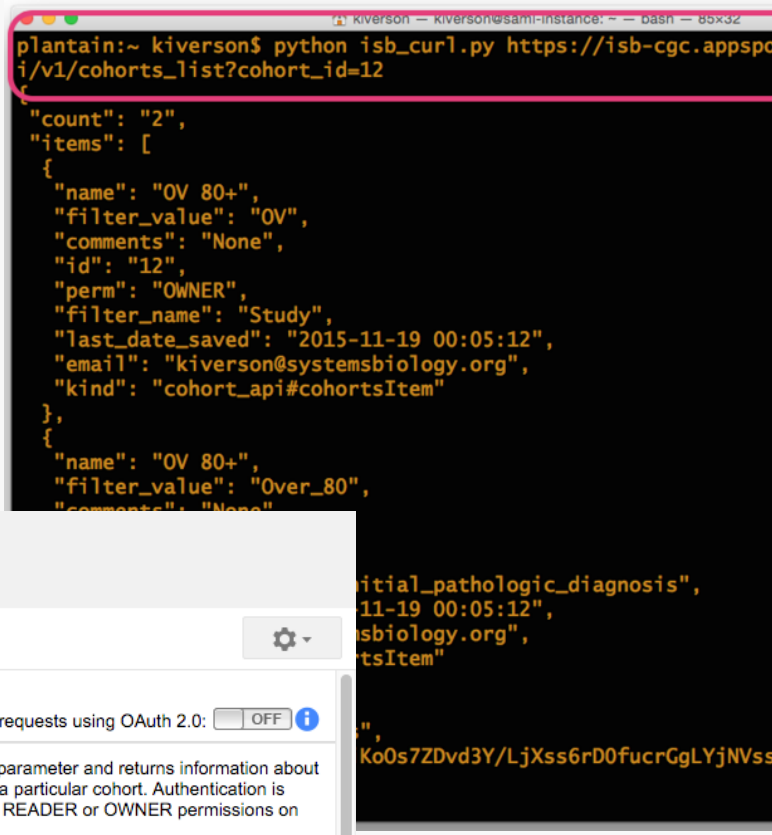
- Cloud Storage
- BigQuery
- Google Genomics
- Compute Engine
- Container Engine
- Cloud Dataflow

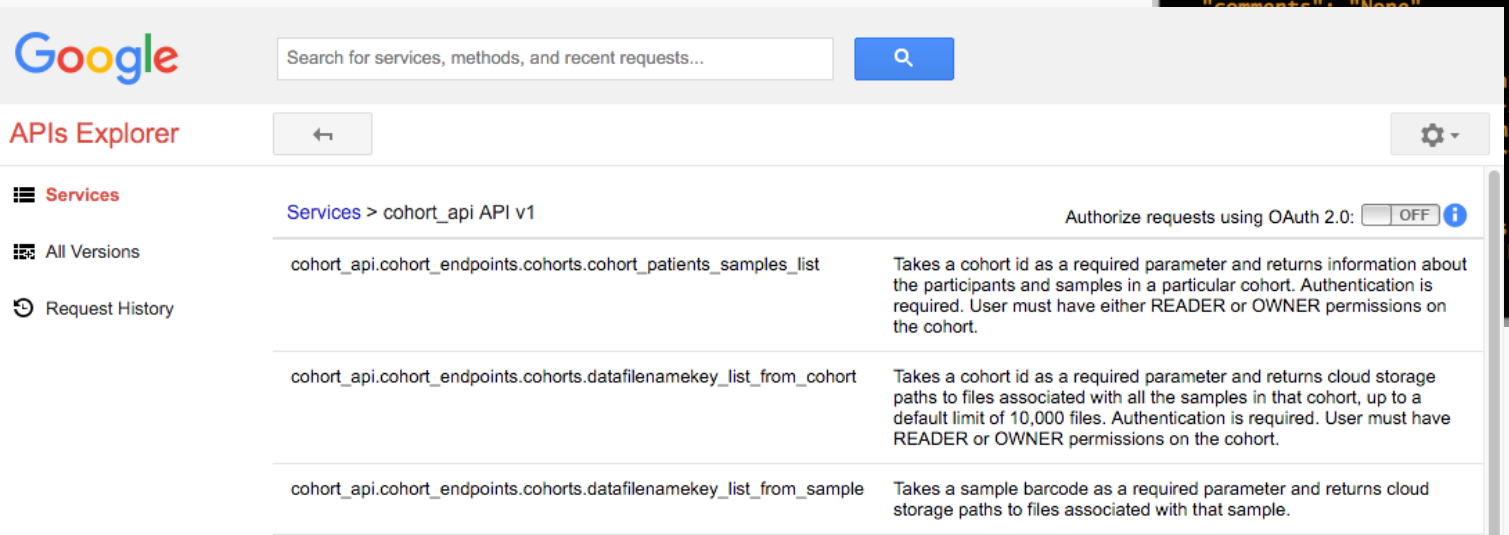


The ISB-CGC API provides programmatic access to the same functionality as the web-app and more:

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





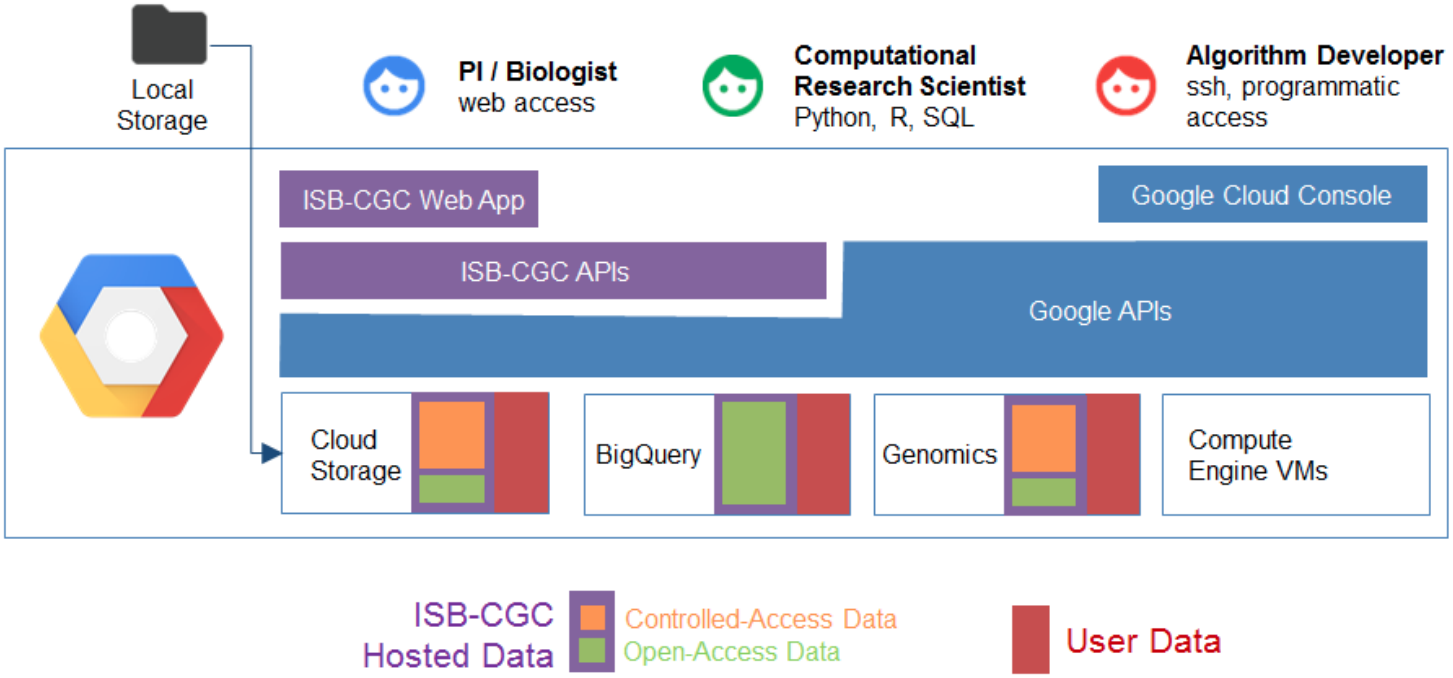
Workflows & Pipelines on VM Clusters

- Grid Engine / Elasticluster
  - use ISB-CGC API combined with Grid Engine running on a Compute Engine (GCE) cluster
- Common Workflow Language (CWL)
  - provision and configure GCE VM for use with CWL
  - create and run CWL workflows
- Kubernetes + GKE (Container Engine)
  - command line interface for launching containerized workflows
- Google Genomics “Pipelines API”
  - released in February
  - Docker-based
  - encapsulates VM-provisioning and data-staging

The ISB Cancer Genomics Cloud

Welcome to the ISB-CGC Documentation on Read the Docs.

Here you will find information describing the features of the ISB-CGC platform, tips on how to use it, and details about the data that we are hosting on the Google Cloud Platform.



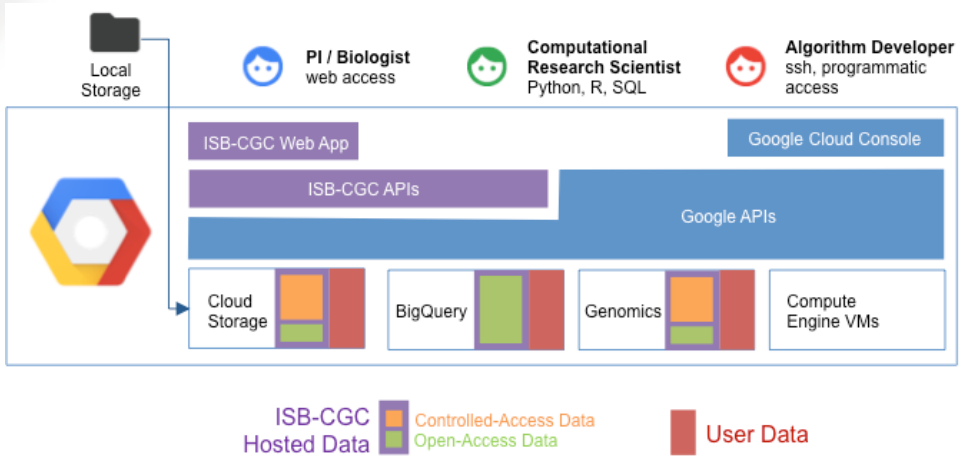
The ISB-CGC aims to serve the needs of a broad range of cancer researchers ranging from scientists or clinicians who prefer to use an interactive web-based application to access and explore the rich TCGA dataset, to computational scientists who want to write their own custom scripts using languages such as R or Python, accessing the data through APIs, and to algorithm developers who wish to spin up thousands of virtual machines to analyze hundreds of terabytes of sequence data.

This documentation is a work-in-progress, please let us know how we can improve it! feedback@isb-cgc.org

– the ISB-CGC team

ISB Cancer Genomics Cloud

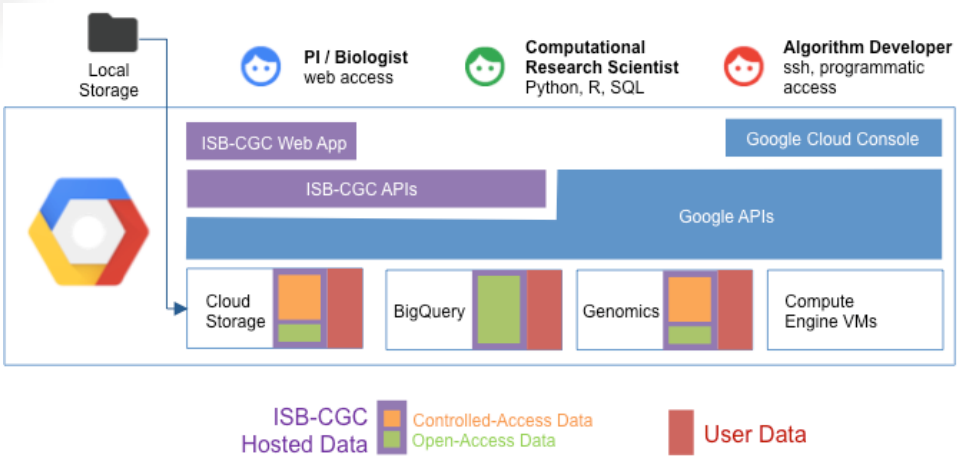
Thank You



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

ISB Cancer Genomics Cloud

Questions?



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