The ISB Cancer Genomics Cloud

Workshop at ISB

August 2nd 2016

ISB-CGC Team Members

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

Jonathan Bingham
Matt Bookman
Nicole Deflaux
Larry Quesada

David Pot
Mark Backus
Ross Casanova
Madelyn Reyes

Our Goals for this Workshop

for YOU

- to understand what the ISB-CGC platform provides
- to know how to find and use the data and tools that suit your needs
- to know your way around the Google Cloud Platform

for US

- to better understand your use-cases, research goals, and needs
- to get feedback & suggestions
    - new data or metadata, sources, and/or formats?
    - new features?
    - new tools?
    - other ideas?

What the ISB-CGC is...

<u>Open</u> platform integrated with the Google Cloud, providing <u>Data</u>, <u>Tools</u>, and <u>Code Samples</u> for a broad range of users

- Data as a Service (DaaS)
  - TCGA clinical and molecular data (multiple formats and technologies)
  - Genome- and Platform-Reference data (*eg* GENCODE, miRTarBase, Kaviar, *etc*)
  - Additional open-access data sets (*eg* CCLE)

- Tools & Applications (SaaS)
  - web app allows interactive and custom visualizations of the TCGA data
  - R and Python examples to get you started implementing your own custom analyses
  - programmatic API endpoints to supplement Google APIs
  - framework to help you use the new Google Pipelines API to run and manage tens of thousands of compute tasks

What the ISB-CGC is…

Open platform integrated with the Google Cloud, providing
Data, Tools, and Code Samples for a broad range of users

- Data as a Service (DaaS)
  - TCGA clinical and molecular data (multiple formats and technologies)
  - Genome- and Platform-Reference data (*eg* GENCODE, miRTarBase, Kaviar, *etc*)
  - Additional open-access data sets (*eg* CCLE)

- Tools & Applications (SaaS)
  - web app allows interactive and custom visualizations of the TCGA data
  - R and Python examples to get you started implementing your own custom analyses
  - programmatic API endpoints to supplement Google APIs
  - framework to help you use the new Google Pipelines API to run and manage tens of thousands of compute tasks
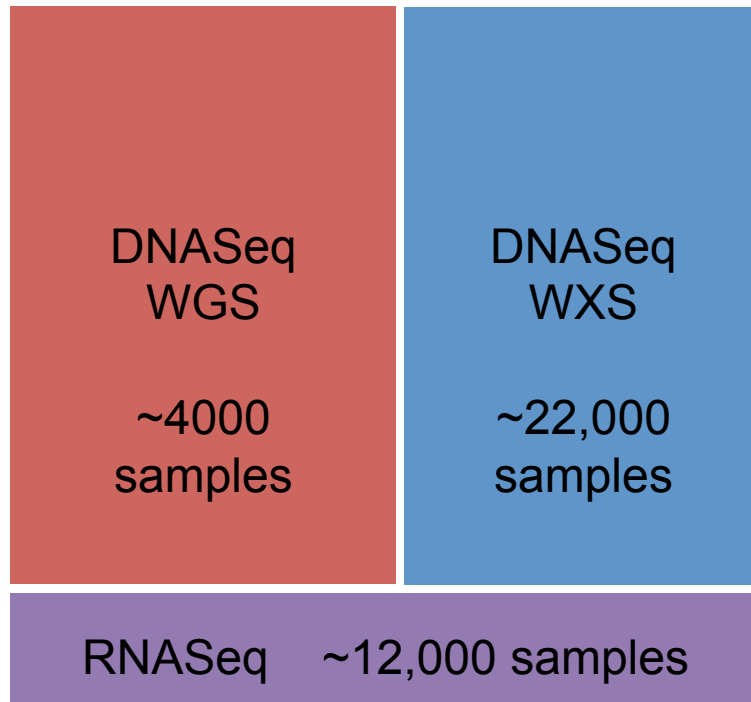
You are the builder

What the ISB-CGC is _not_ ...

- a repository from which to _download_ large data sets
    - bring _your_ tools and data to the cloud, to work with _our_ tools and data!

- a "walled garden" with a _single_ entrance
    - only "fence" is around the controlled-access data, but once you have a key, you can choose the path you prefer (GUI, CLI, API)

- a repository of "approved" or "recommended" pipelines for performing standard tasks
    - best left to the research community (_eg_ DREAM challenges, GA4GH, PCAWG, tools developers, and you!)

TCGA Size & Complexity

>1 PB of sequence data
(controlled access)

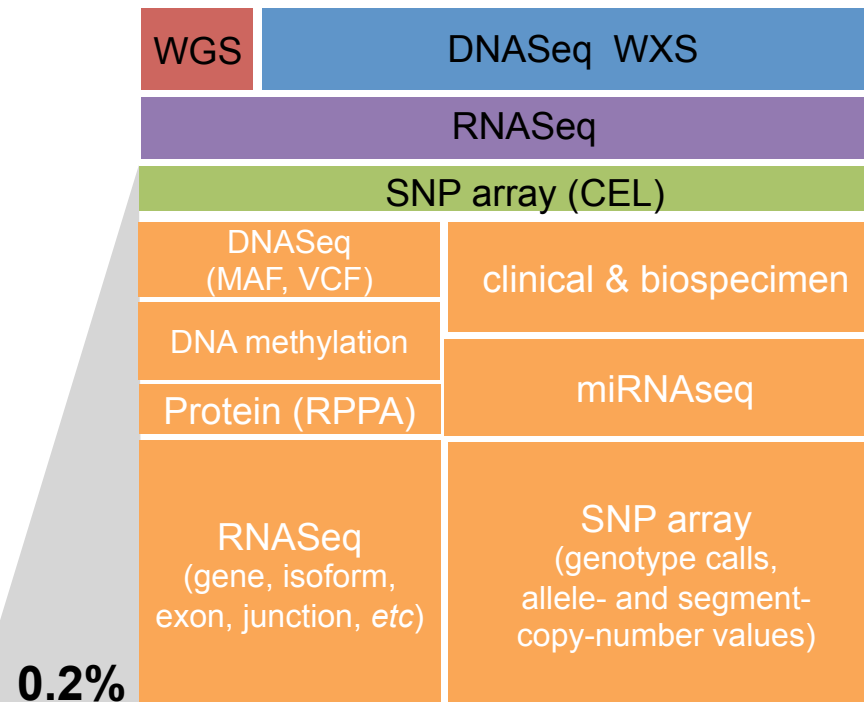~400,000 files of
heterogeneous data
(mostly open-access)

DNASeq WGS ~4000 samples

DNASeq WXS ~22,000 samples

RNASeq ~12,000 samples

WGS

DNASeq WXS

RNASeq

SNP array (CEL)

DNASeq (MAF, VCF)

DNA methylation

Protein (RPPA)

clinical & biospecimen

miRNAseq

RNASeq (gene, isoform, exon, junction, *etc*)

SNP array (genotype calls, allele- and segment-copy-number values)

0.2%

Table Details: Clinical_data

Schema

Field	Type	Nullable	
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	
clinical_stage	STRING		
colorectal_cancer			
country			
vital_status			
days_to_birth			
days_to_death			
days_to_last_known_alive			
days_to_last_followup			
days_to_initial_pathologic_diagnosis			
days_to_submitted_specimen_dx			
ethnicity			
frozen_specimen_anatomic_site			
gender			
gleason_score_combined			
histological_type			
history_of_colon_polyps			

Table Details: Biospecimen_data

Schema

Field	Type	Nullable	
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	
max_percent_lymphocyte_infiltrat			
max_percent_monocyte_infiltrati			
max_percent_necrosis			
max_percent_neutrophil_infiltrati			
max_percent_normal_cells			
max_percent_stromal_cells			
max_percent_tumor_cells			
max_percent_tumor_nuclei			

Table Details: Annotations

Schema

Field	Type	Nullable	
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: Protein_RPPA_data

Schema

Field	Type	Nullable	
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: Somatic_Mutation_calls

Schema

Field	Type	Nullable	
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	
GENCODE_Transcript_Name	S		
GENCODE_Transcript_Status	S		
GENCODE_Transcript_Type	S		
GO_Biological_Process	S		
GO_Cellular_Component	S		
GO_Molecular_Function	S		
Gene_Type	S		
Genome_Change	S		

Table Details: Copy_Number_segments

Schema

Field	Type	Nullable	
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: mRNA_UNC_HiSeq_RSEM

Schema

| Field	Type	Nullable	
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	NULLABLE	Describe this field...
Platform	STRING	NULLABLE	Describe this field...
original_gene_symbol	STRING	NULLABLE	Describe this field...
HGNC_gene_symbol	STRING	NULLABLE	Describe this field...
gene_id	INTEGER	NULLABLE	Describe this field...
normalized_count	FLOAT	NULLABLE	Describe this field...

Table Details: mRNA_BCGSC_HiSeq_RPKM

Schema

Field	Type	Nullable	
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...
	INTEGER	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...
	FLOAT	NULLABLE	Describe this field...
	STRING	NULLABLE	Describe this field...

Table Details: DNA_Methylation_betas

Schema

Field	Type	Nullable	
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 methylat

Table Details: miRNA_expression

Schema

| Field	Type	Nullable	
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...

ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!
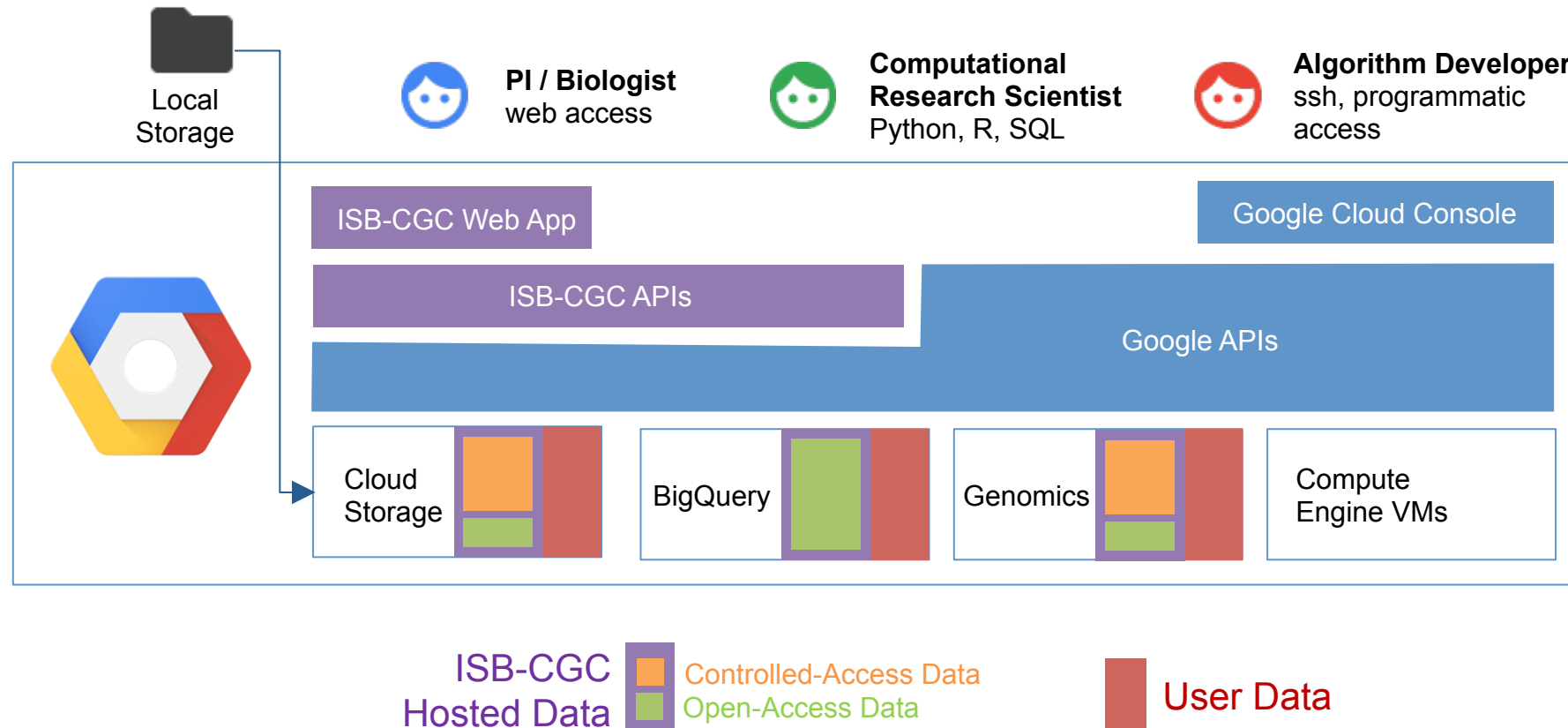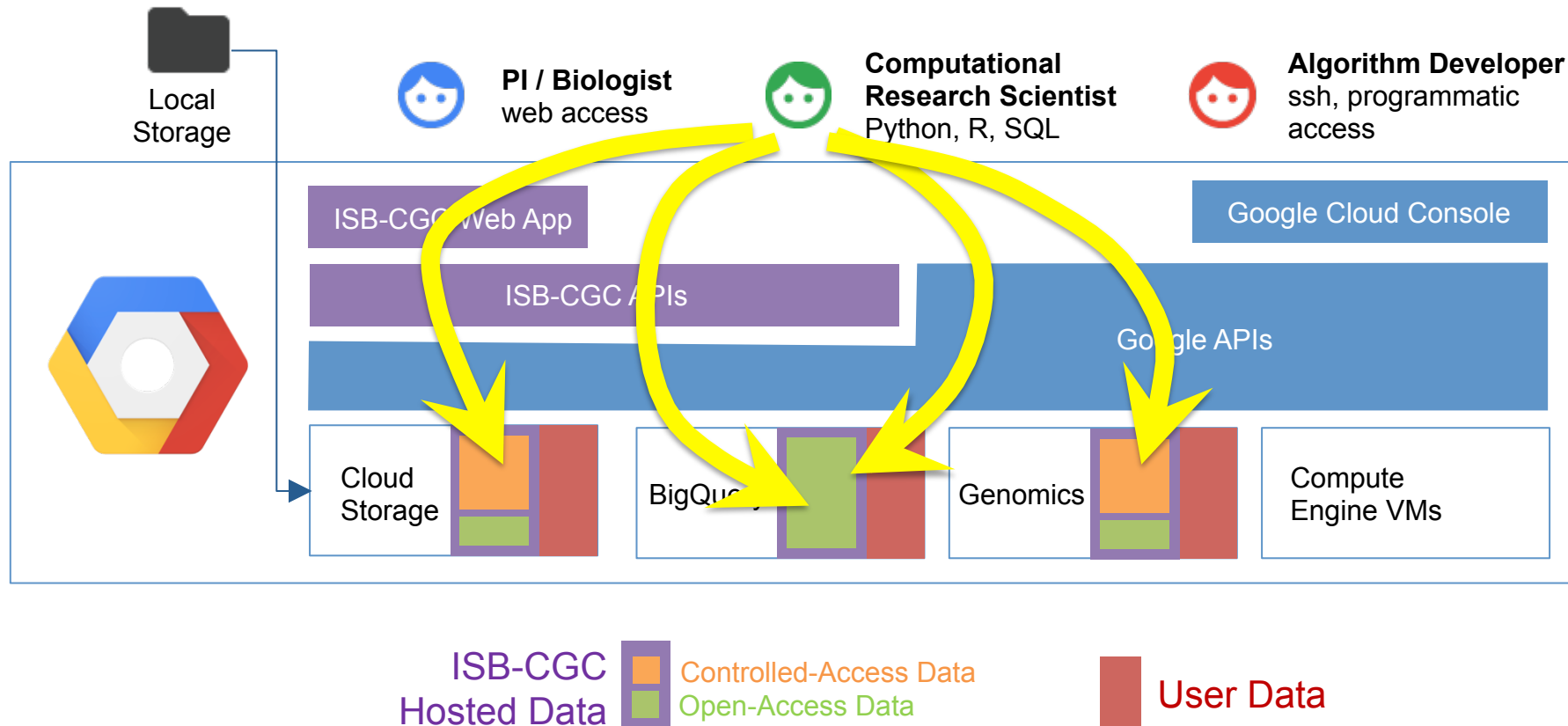
ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!
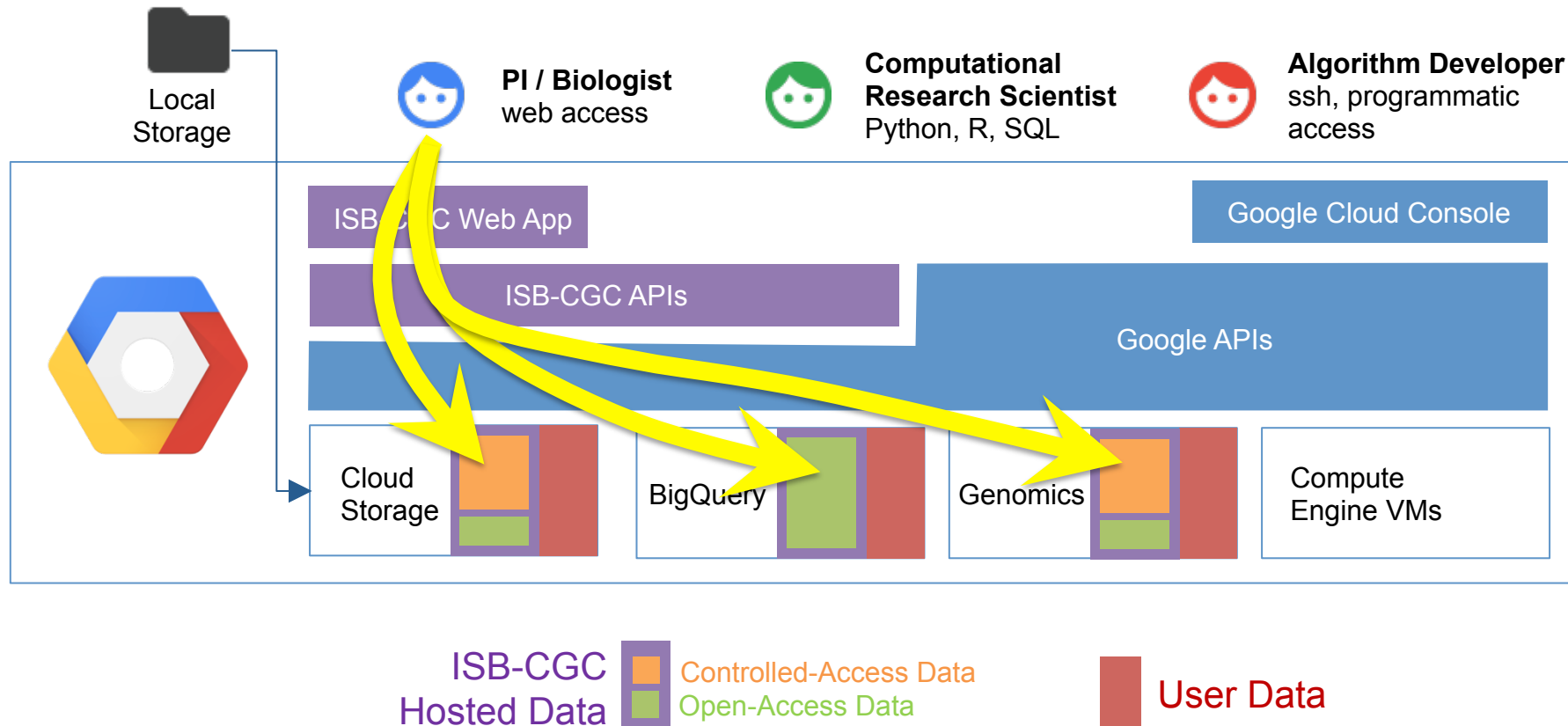
Web access for the PI / Biologist:
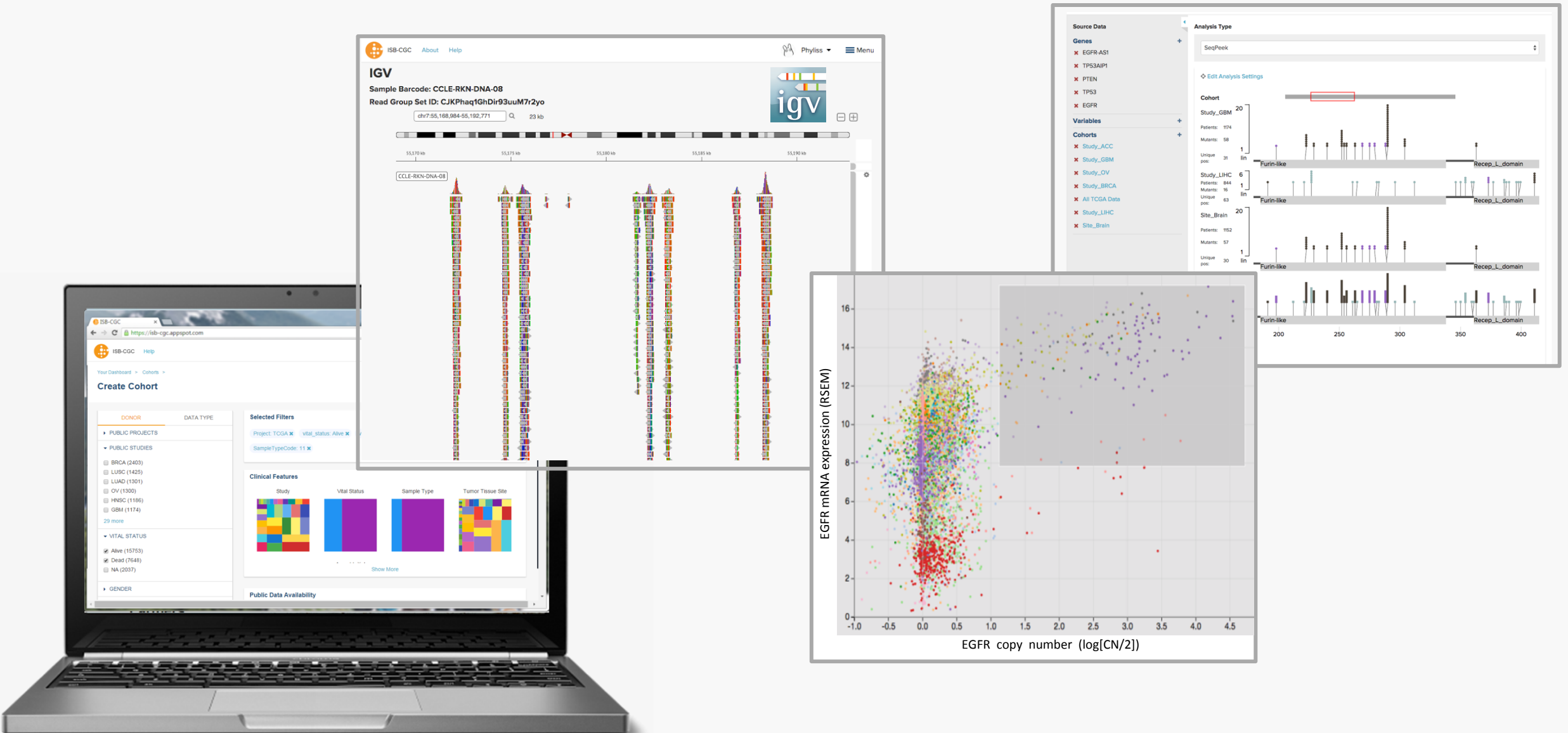
ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!

Python, R, and SQL for the Computational Scientist:

ISB-CGC Approach

- build an open platform that can grow and evolve to satisfy a broad range of users and use-cases
- leverage the best existing tools and technologies, as they are released
- collaborate with the research community in areas of data standards, containers and workflows, *etc*
- provide a range of examples and tutorials to get newcomers up and running quickly!

A generalizable workflow using Docker