CANCER GENOMICS CLOUD
SEVEN BRIDGES

Scaling Bioinformatics on the Cancer Genomics Cloud

Dave Roberson
Community Engagement
October 15, 2019
BYOB - NIH

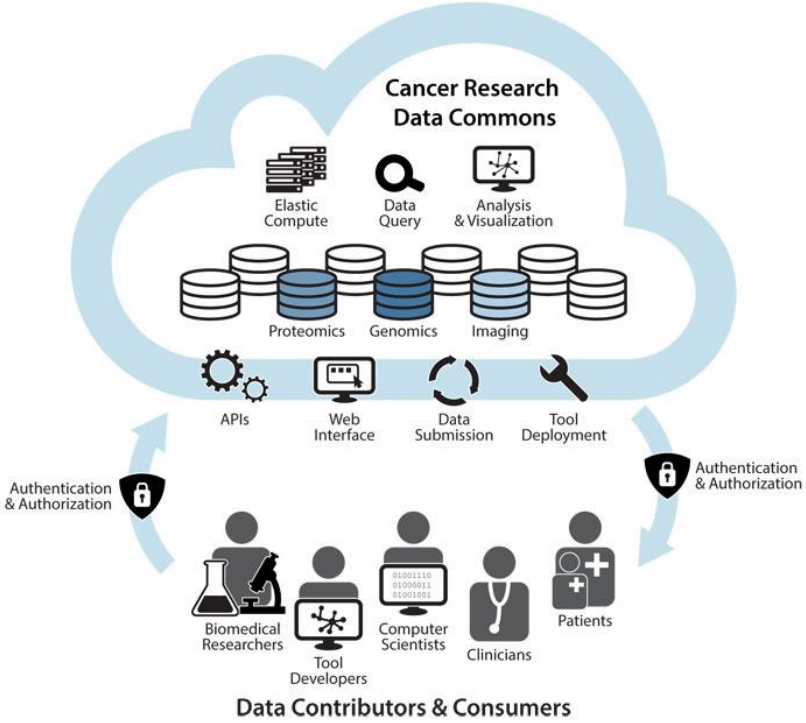
Community Feedback Needed

- Message us on BYOB slack
- Survey distributed to the email list
- Email us directly at cgc@sevenbridges.com

Agenda

1. Motivations for Biomedical Analysis Platforms
2. Scaling large cohort generation and QC for association studies
3. How you can get involved

The Seven Bridges Cancer Genomics Cloud (CGC)



A Cloud Resource within the NCI Cancer Research Data Commons for secure storage, sharing & analysis of petabytes of public, multi-omic cancer datasets

NATIONAL CANCER INSTITUTE

The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E.

Analysis Platforms Enable:

1. International collaborations and consortia
2. Educational resources and community standards
3. A network of FAIR data
    - *Findable Accessible Interoperable Reusable*
4. Secure workspaces/sandboxes
5. Multi-cloud computation (AWS and Google)

Precision Medicine Ecosystem

Infrastructure

PDXNet

CAVATICA

DataSTAGE
Powered by Seven Bridges

Interoperability

COMMON WORKFLOW LANGUAGE

Rabix
[Reproducible Analysis for Bioinformatics]

BioCompute Objects

Global Alliance for Genomics & Health

Partnerships

NATIONAL CANCER INSTITUTE

Genomics England

GEORGETOWN UNIVERSITY

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

TCGA

PCAWG
PanCancer Analysis of WHOLE GENOMES

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Datasets accessible from the CGC

Dataset	Description	Experimental setup	File types
TCGA	Rich dataset of 11 tumor types and 7 experimental strategies	WES, RNAseq, miRNAseq, methylation, genotyping, ATACseq, imaging	BAM
TARGET	Dataset of genomic changes in childhood cancers	RNASeq, WGS, WES, miRNAseq	BAM, MAF, TSV, VCF
CANCER IMAGING ARCHIVE	Imaging data from many 21 tumor types	Imaging	DCM
CPTAC	Proteomics of 10 tumor types and associated genomic data	Proteomics, WGS, WES, RNAseq	BAM, TSV, VCF, mzML.gz, mzid.gz, raw, tar.gz
International Cancer Genome Consortium	Consortium of many datasets, 20 studies on CGC	WGS, RNASeq	BAM, VCF
CCLE Cancer Cell Line Encyclopedia	Dataset of 1457 cancer cell lines	WGS, WES, RNAseq	BAM
SIMONS FOUNDATION	Genome sequencing of 130 populations	WGS	BAM, VCF
Personal Genome Project	Crowdsourced genomics, datasets from 10 individuals	WGS, WGBS, RNAseq, methylation	BAM, FASTQ, IDAT, TBI, VCF
HUMAN CELL ATLAS	Single-cell genomics of healthy tissues	RNASeq	FASTQ

Docker and CWL Enables Reproducible Analysis

Docker Container

Easily share binaries, libraries, and dependencies for your applications

Common Workflow Language

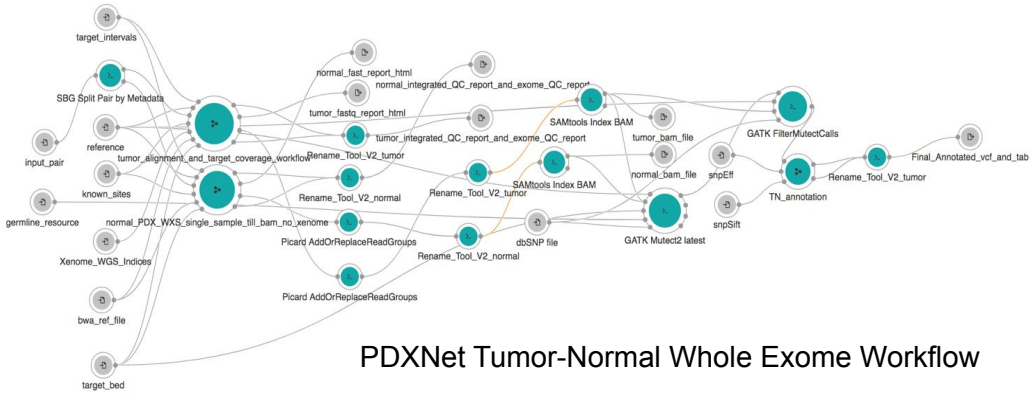
Standardized YAML description of tools and workflows (e.g. commands, parameters)

Rabix Suite

Composer: an integrated development environment to code, test, and debug CWL apps
Executor: run CWL apps locally or at scale on HPC or cloud environments

SevenBridges

Sync your apps to Seven Bridges platforms to analyze your data alongside massive public datasets like TCGA



PDXNet Tumor-Normal Whole Exome Workflow

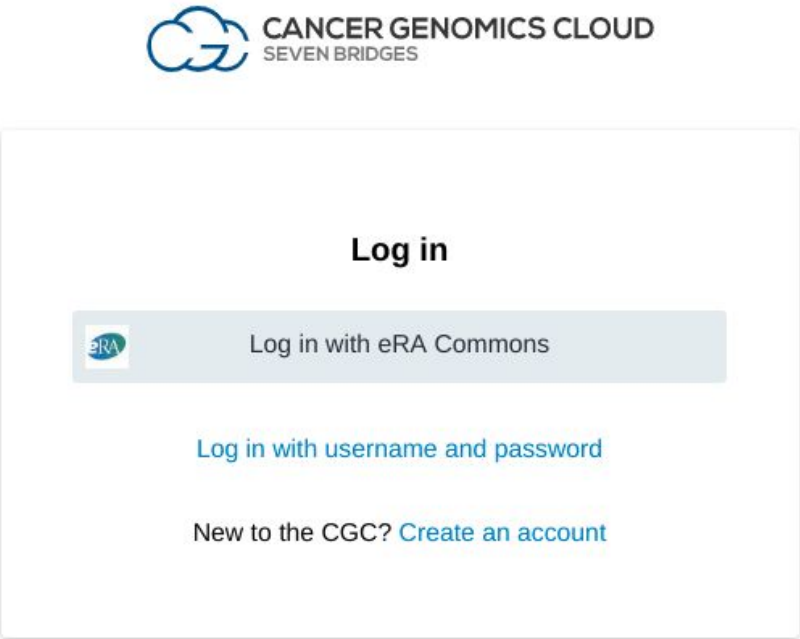
Wide range of research enabled by the CGC

4,000$^+$ users from **80$^+$** countries have used the CGC to run **980,000$^+$** computational tasks representing **1000$^+$** years of total compute time to:

- Detect aberrant splice junctions and splicing profiles across patient populations
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across patient populations
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across patient populations
- Detect novel gene fusions from RNA-Seq data
- Identify cis-regulatory region variants across patient populations
- ...and much more

Very Easy To Get Started

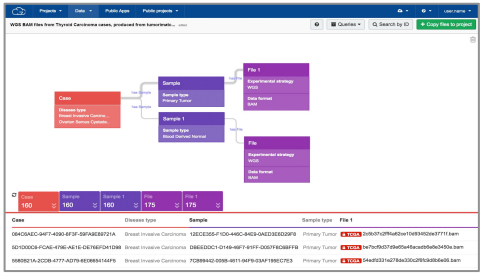
- Free to sign up cgc.sbgenomics.com
- Option to connect with eRA Commons to access controlled data
- $300 of pilot funding to get your project started
- Comprehensive online documentation and training resources
- Technical support from a team of scientists, bioinformaticians, and engineers

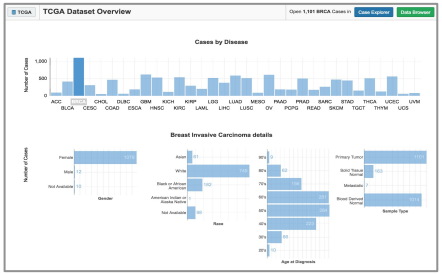


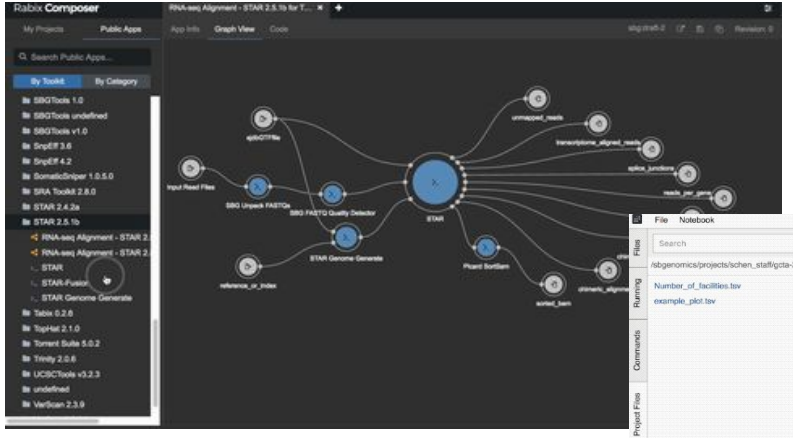
CGC provides an easy way to find and analyze data

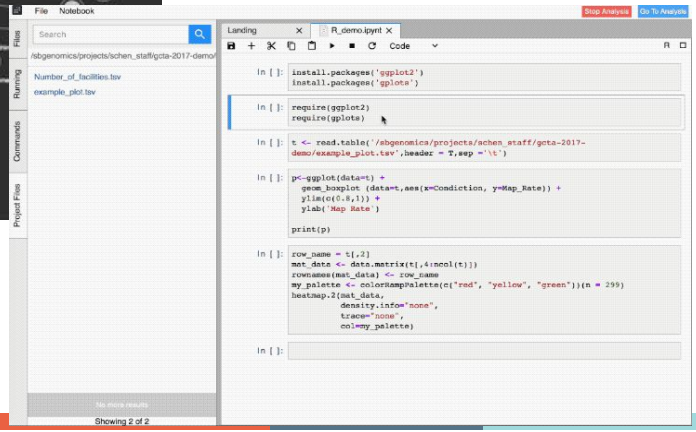
Visually explore and access **3+ PB** of multi-omic public data through interactive query tools & APIs.

Use the **400+** cloud- and cost-optimized tools in our Public Apps library OR deploy custom tools using **Rabix Composer**, Jupyter notebooks or R packages



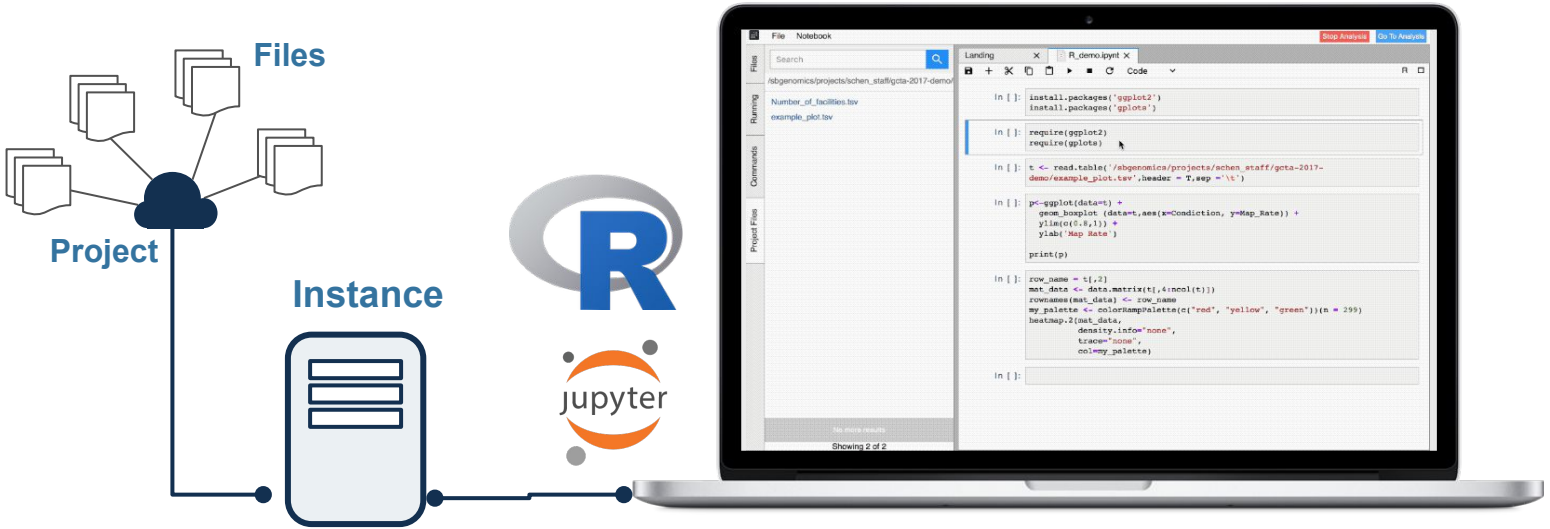






Powerful, collaborative, & reproducible interactive analysis

Users create interactive analysis sessions within a project - all files are available and over 50 instances can be used (*c3.2xlarge* to *x1.32xlarge* on AWS)



Demo: Cohort QC At Scale Using HAIL and Other Tools

Projects Data Public Apps Public projects Developer Staff jack_digi

New query

Select dataset(s)

Explore selected

Featured datasets

Featured datasets to explore Learn More

TCGA GRCh38 GDC v16	11,315 Case	27 DiseaseType	314,354 File
TARGET GRCh38 GDC v16	3,367 Case	34 DiseaseType	9,640 File
CPTAC	335 Case	14,114 File	7 Protocol
TCIA	1,777 Case	21,265 File	3,216 Study
ICGC	1,432 Donor	35,766 File	20 Project

Forum Terms Privacy Data Use

© 2019 Seven Bridges Genomics

Projects | Data | Public Apps | Public projects | Automations | Developer | Staff | david.roberson

Projects — byob_demo

byob_demo
Created by david.roberson · Oct. 13, 2019 21:16

+ Create a project | View all projects

Public Data and Apps

Analyze
549,625
Overview | Cases | Browse Data

Use some of
200
publicly available
Tools and Workflows
Browse Apps

Analyses — GATK4 Joint Discovery

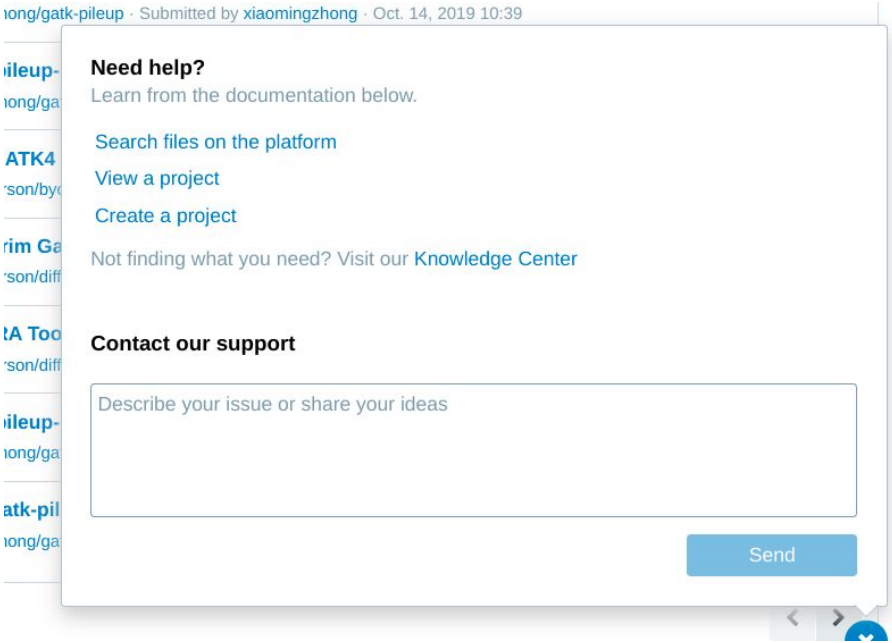
Tasks | Data Cruncher

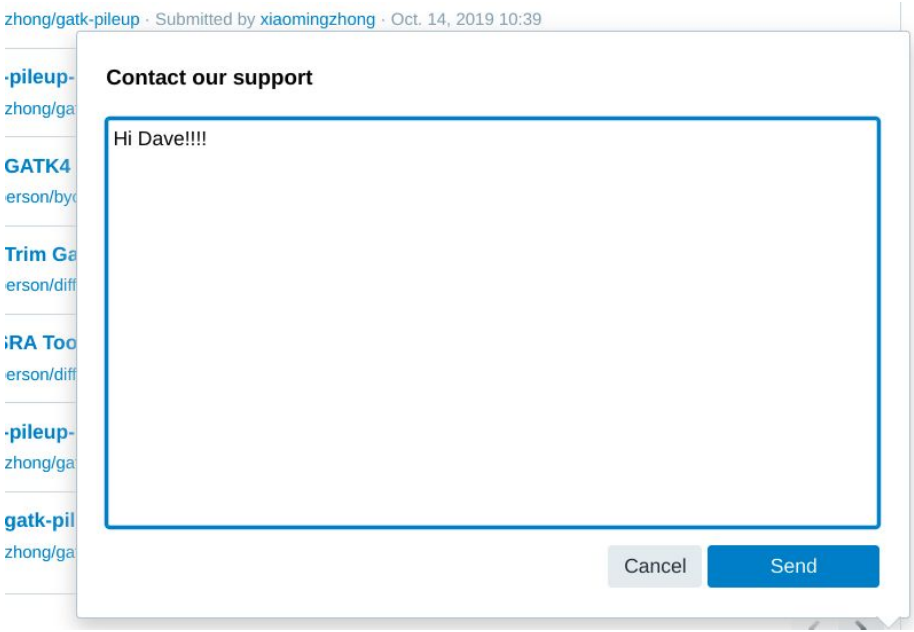
COMPLETED **GATK4 Joint Discovery run - 10-14-19 01:20:31**
Project: david.roberson/byob-demo · Submitted by david.roberson · Oct. 13, 2019 21:36

COMPLETED **GATK4 Joint Discovery run - 06-23-19 12:49:42**
Project: jack_digi/international-lymphoma-epidemiology-consortium · Submitted by jack_digi · June 23, 2019 08:53

COMPLETED **GATK4 Joint Discovery run - Demo Ashkenazim**
Project: anaDsbg/vijai-conference-demo · Submitted by anaDsbg · June 17, 2019 17:31

Forum | Terms | Privacy | Data Use


ong/gatk-pileup · Submitted by xiaomingzhong · Oct. 14, 2019 10:39
Need help?
Learn from the documentation below.
Search files on the platform
View a project
Create a project
Not finding what you need? Visit our Knowledge Center
Contact our support
Describe your issue or share your ideas
Send


zhong/gatk-pileup · Submitted by xiaomingzhong · Oct. 14, 2019 10:39
Contact our support
Hi Dave!!!!
Cancel
Send

Projects | Data | Public Apps | Public projects | Automations | Developer | Staff | david.roberson

Dashboard | Files | Apps | Tasks

byob_demo

Interactive Analysis | Settings | Notes

Description

This is markdown

Edit description

Members

Email notifications

david.roberson OWNER
Write, Copy, Execute, Admin

Don't work alone.
The best research happens in teams.

Invite new members

Share your tools, data, and ideas with collaborators

Analyses

Search

Tasks | Data Cruncher

COMPLETED **GATK4 Joint Discovery run - 10-14-19 01:20:31**
Submitted by david.roberson · Oct. 13, 2019 21:36

COMPLETED **GATK4 Joint Discovery run - 10-14-19 01:20:31**

Get support | View stats & logs | Edit and rerun

Executed on Oct. 13, 2019 21:36 by david.roberson

Spot Instances: **On** | Memoization: **Off** | Price: **$0.87** Refund View refunds | Duration: **1 hour, 22 minutes**

App: GATK4 Joint Discovery - Revision: 1

Inputs

axiomPoly_resource_vcf
- Axiom_Exome_Plus.genotypes.all_populations.poly.hg38.vcf.gz

dbsnp_vcf
- Homo_sapiens_assembly38.dbsnp138.vcf

hapmap_resource_vcf
- hapmap_3.3.hg38.vcf.gz

input_vcfs
- HG002-NA24385-50x.50.merged.g.vcf.gz
- HG003.hs37d5.60x.1.converted.50.merged.g.vcf.gz
- HG004.hs37d5.60x.1.converted.50.merged.g.vcf.gz

intervals
- intervals_hg38.even.handcurated.20k.intervals

mills_resource_vcf
- Mills_and_1000G_gold_standard.indels.hg38.vcf.gz

omni_resource_vcf
- 1000G_omni2.5.hg38.vcf.gz

one_thousand_genomes_resource_vcf
- 1000G_phase1.snps.high_confidence.hg38.vcf.gz

reference_fasta

App Settings

Show non-default

GatherVcfs (#gatk_gathervcfs_1)

out_prefix (Output prefix)	ashkenazim

CollectVariantCallingMetrics (#collectvariantcallingmetrics)

out_prefix (#output_basename)	ashkenazim

Outputs

Joint VCF
- ashkenazim.vcf.gz

Metrics
- ashkenazim.variant_calling_detail_metrics
- ashkenazim.variant_calling_summary_metrics

Interactive analysis on the Cancer Genomics Cloud

hail for genetic association studies

In this demo, we will work with Hail Python package. Hail enables scalable downstream analysis, mostly in the area of genetic association studies. That is, looking for statistical association between the variants and inferring their connection to disease.

Let's get started!

In [1]:
```
import hail as hl
hl.init()
```

```
using hail jar at /opt/conda/lib/python3.6/site-packages/hail/hail-all-spark.jar
Running on Apache Spark version 2.2.0
SparkUI available at http://172.17.0.2:4040
Welcome to
     __  __     <>__
    / /_/ /__  __/ /
   / __  / _ `/ / /
  /_/ /_/\_,_/_/_/   version 0.2.11-cf54f08305d1
LOGGING: writing to /sbgenomics/workspace/hail-20191014-2149-0.2.11-cf54f08305d1.log
```

In [2]:
```
from hail.plot import show
from pprint import pprint
hl.plot.output_notebook()
```

Loading BokehJS ...

tutorial using 1000g data

Steps 1. through 4. are reproduced from Hail Tutorial available here.

use hail for your data

We will follow a similar flow as the tutorial, but now add in the vcf we calculated using joint-calling.

Add files from your *Project* ¶

DataSTAGE(SB) users can easily add any files within their project to *Data Cruncher* for interactive analysis. To add any file, go to the **Project Files** on the left panel, locate the file, and click on it.

This will automatically copy the relevant path (here `/sbgenomics/project-files/ashkenasi_trio/Demo_Ashkenazim.vcf.bgz`, paste it into the cell below.

Additionally, Ashkenazim VCF file was obtained using GRCh38 reference genome, hence the additional "reference_genome" argument in the following line of code.

DEMO NOTE: Ashkenazim VCF takes about ~50sec to import

In [16]:
```
hl.import_vcf('/sbgenomics/project-files/ashkenasi_trio/Demo_Ashkenazim.vcf.bgz', reference_genome =
'GRCh38').write('Demo_Ashkenazim.mt', overwrite=True)
```

```
2019-09-05 16:22:03 Hail: INFO: Coerced sorted dataset
2019-09-05 16:22:35 Hail: INFO: wrote matrix table with 6684118 rows and 3 columns in 22 partitions to Demo_Ashkenazim.mt
```

To enable **much faster** downstream analysis, we will create a *MatrixTable* from the input VCF

In [17]:
```
ashkenazim = hl.read_matrix_table('Demo_Ashkenazim.mt')
```

Data exploration

In [18]:
```
ashkenazim.rows().select().show(5)
```

Summary: The CGC Enables Efficient and Collaborative Science

- ✓ Quickly identify cohorts
- ✓ Build pipelines from public apps
- ✓ Bring in your own tools and notebooks
- ✓ Share and collaborate securely and easily

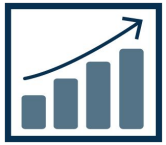
Collaborative Project program to advance your research

- Submit a proposal for up to **$10,000** in cloud credits to **cgc@sevenbridges.com**
- Get additional access to our CGC team and bioinformatics support
- Projects have resulted in dozens of papers, many users submit multiple papers from one project

Seven Bridges Cancer Genomics Cloud - sign up today!

Register for a free account today at http://cgc.sbgenomics.com/
Questions? Contact the Seven Bridges CGC Team at cgc@sevenbridges.com

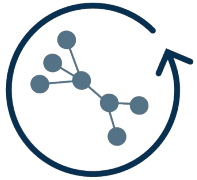
Easy data management



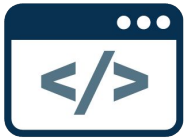
Scalable computation

Secure collaboration

Optimized bioinformatics algorithms



Flexible & fully reproducible methods



Extensible and developer-friendly platform