Running Workflows in the Cloud

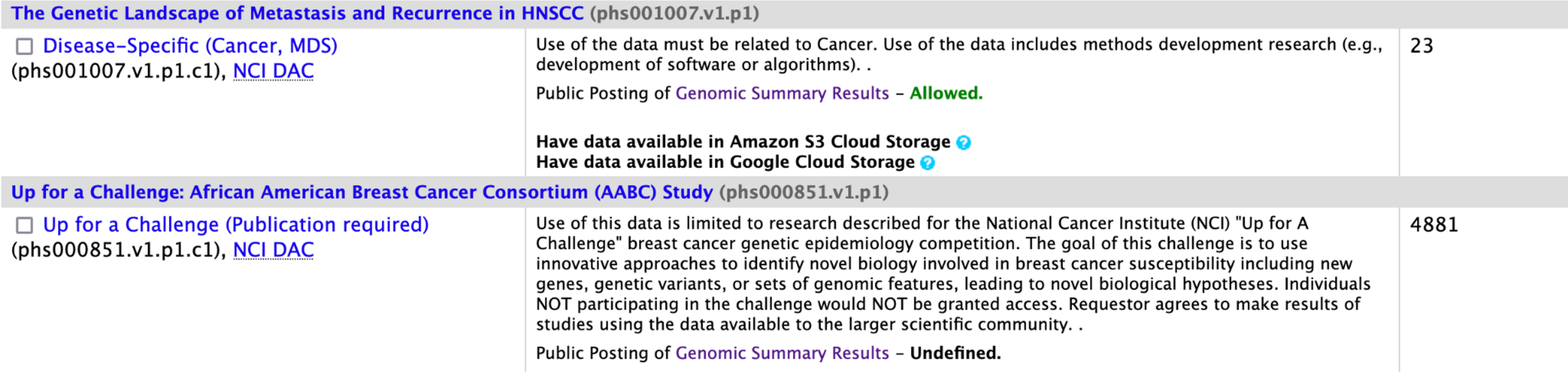
BFX Workshop
Jason Walker, Chris Miller

Getting data from dbGaP

1) Apply for access

- ERA Commons account

- Fill out a Data Access Request (DAR)
  - How are you going to use the data? Read the study information



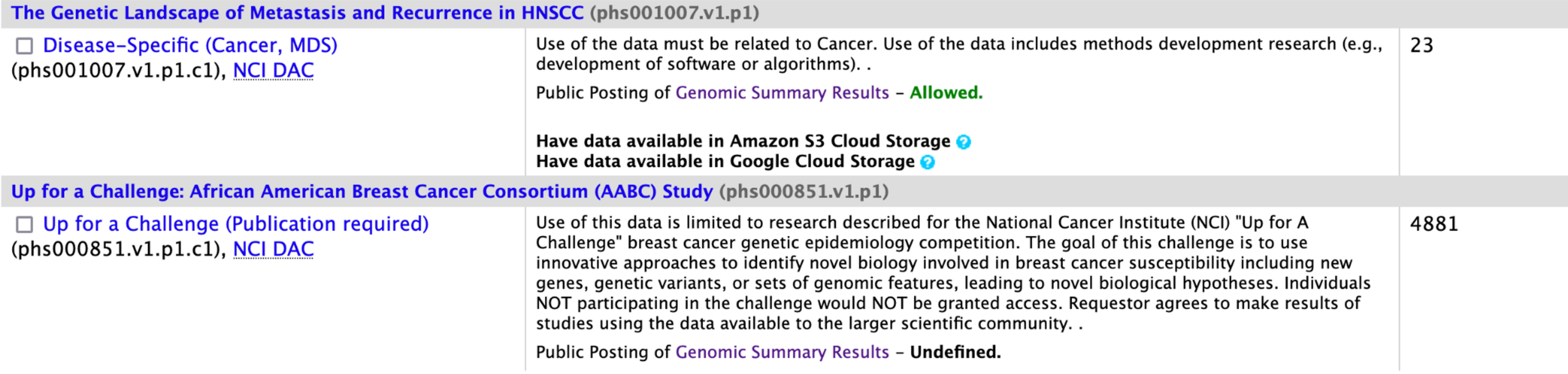
The Genetic Landscape of Metastasis and Recurrence in HNSCC (phs001007.v1.p1)		
□ Disease-Specific (Cancer, MDS) (phs001007.v1.p1.c1), NCI DAC	Use of the data must be related to Cancer. Use of the data includes methods development research (e.g., development of software or algorithms). . Public Posting of Genomic Summary Results - **Allowed.** **Have data available in Amazon S3 Cloud Storage** ? **Have data available in Google Cloud Storage** ?	23
Up for a Challenge: African American Breast Cancer Consortium (AABC) Study (phs000851.v1.p1)		
□ Up for a Challenge (Publication required) (phs000851.v1.p1.c1), NCI DAC	Use of this data is limited to research described for the National Cancer Institute (NCI) "Up for A Challenge" breast cancer genetic epidemiology competition. The goal of this challenge is to use innovative approaches to identify novel biology involved in breast cancer susceptibility including new genes, genetic variants, or sets of genomic features, leading to novel biological hypotheses. Individuals NOT participating in the challenge would NOT be granted access. Requestor agrees to make results of studies using the data available to the larger scientific community. . Public Posting of Genomic Summary Results - **Undefined.**	4881

Getting data from dbGaP

1) Apply for access

- ERA Commons account

- Fill out a Data Access Request (DAR)
    - How are you going to use the data? Read the study information



The Genetic Landscape of Metastasis and Recurrence in HNSCC (phs001007.v1.p1)		
□ Disease-Specific (Cancer, MDS) (phs001007.v1.p1.c1), NCI DAC	Use of the data must be related to Cancer. Use of the data includes methods development research (e.g., development of software or algorithms). . Public Posting of Genomic Summary Results – **Allowed.** **Have data available in Amazon S3 Cloud Storage** **Have data available in Google Cloud Storage**	23
Up for a Challenge: African American Breast Cancer Consortium (AABC) Study (phs000851.v1.p1)		
□ Up for a Challenge (Publication required) (phs000851.v1.p1.c1), NCI DAC	Use of this data is limited to research described for the National Cancer Institute (NCI) "Up for A Challenge" breast cancer genetic epidemiology competition. The goal of this challenge is to use innovative approaches to identify novel biology involved in breast cancer susceptibility including new genes, genetic variants, or sets of genomic features, leading to novel biological hypotheses. Individuals NOT participating in the challenge would NOT be granted access. Requestor agrees to make results of studies using the data available to the larger scientific community. . Public Posting of Genomic Summary Results – **Undefined.**	4881

Getting data from dbGaP

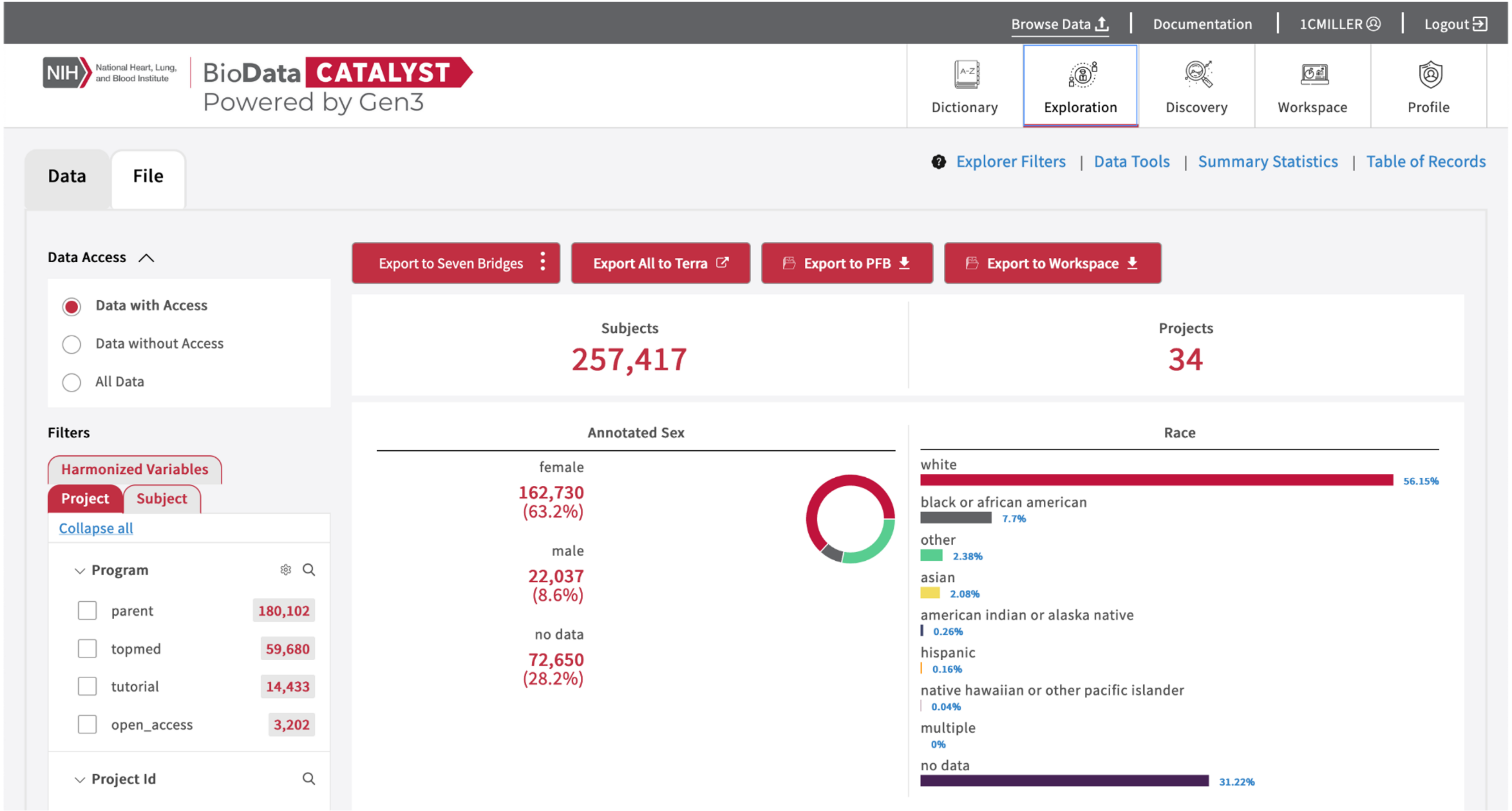
1) Apply for access

- ERA Commons account

- Fill out a Data Access Request (DAR)
  - How are you going to use the data? Read the study information

- Wait 2-8 weeks for approval

https://gen3.biodatacatalyst.nhlbi.nih.gov/



Data Access

- ● Data with Access
- ○ Data without Access
- ○ All Data

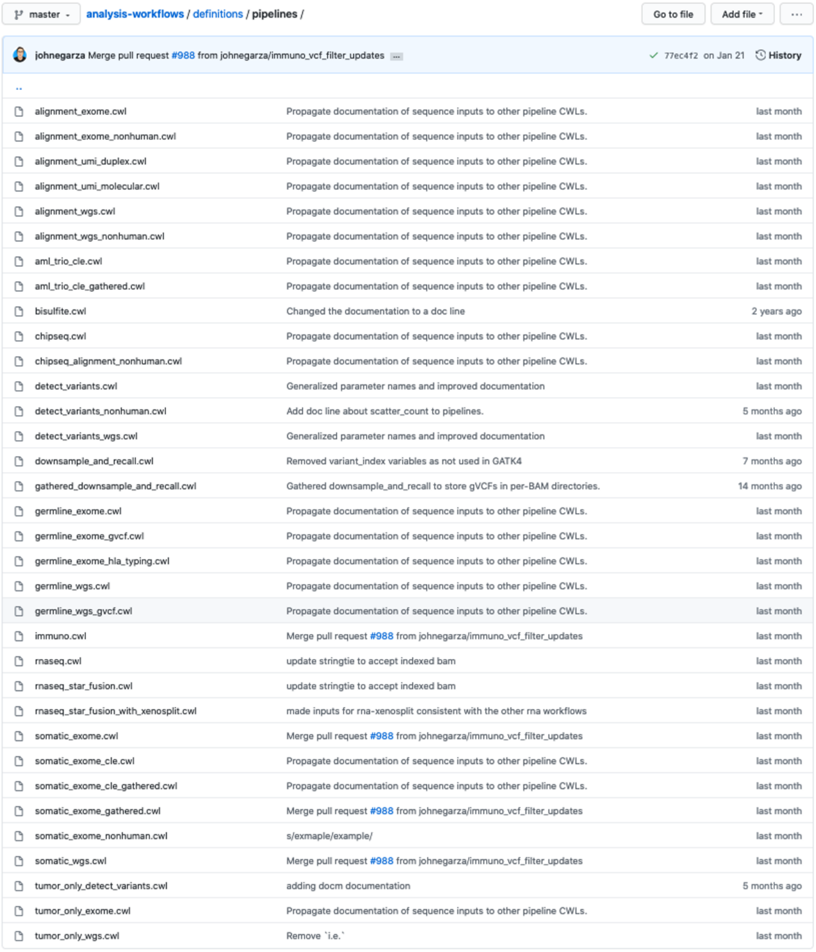
Filters

Program

□ parent	180,102
□ topmed	59,680
□ tutorial	14,433
□ open_access	3,202

- 400,000 lines of code in the `genome/analysis-workflows` github repository

- Dozens of data types and approaches
  - exome/WGS/targeted (somatic/germline)
  - bisulfite
  - RNAseq
  - single-cell (TCR, 5’/3’, ATAC)
  - RNAseq (expression, fusions, splicing)
  - ATAC/ChIPseq
  - etc



johnegarza Merge pull request #988 from johnegarza/immuno_vcf_filter_updates … ✓ 77ec4f2 on Jan 21

..		
alignment_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_exome_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_umi_duplex.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_umi_molecular.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_wgs.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_wgs_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
aml_trio_cle.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
aml_trio_cle_gathered.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
bisulfite.cwl	Changed the documentation to a doc line	2 years ago
chipseq.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
chipseq_alignment_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
detect_variants.cwl	Generalized parameter names and improved documentation	last month
detect_variants_nonhuman.cwl	Add doc line about scatter_count to pipelines.	5 months ago
detect_variants_wgs.cwl	Generalized parameter names and improved documentation	last month
downsample_and_recall.cwl	Removed variant_index variables as not used in GATK4	7 months ago
gathered_downsample_and_recall.cwl	Gathered downsample_and_recall to store gVCFs in per-BAM directories.	14 months ago
germline_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_exome_gvcf.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_exome_hla_typing.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_wgs.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_wgs_gvcf.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
immuno.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
rnaseq.cwl	update stringtie to accept indexed bam	last month
rnaseq_star_fusion.cwl	update stringtie to accept indexed bam	last month
rnaseq_star_fusion_with_xenosplit.cwl	made inputs for rna-xenosplit consistent with the other rna workflows	last month
somatic_exome.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
somatic_exome_cle.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
somatic_exome_cle_gathered.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
somatic_exome_gathered.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
somatic_exome_nonhuman.cwl	s/exmaple/example/	last month
somatic_wgs.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
tumor_only_detect_variants.cwl	adding docm documentation	5 months ago
tumor_only_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
tumor_only_wgs.cwl	Remove `i.e.`	last month

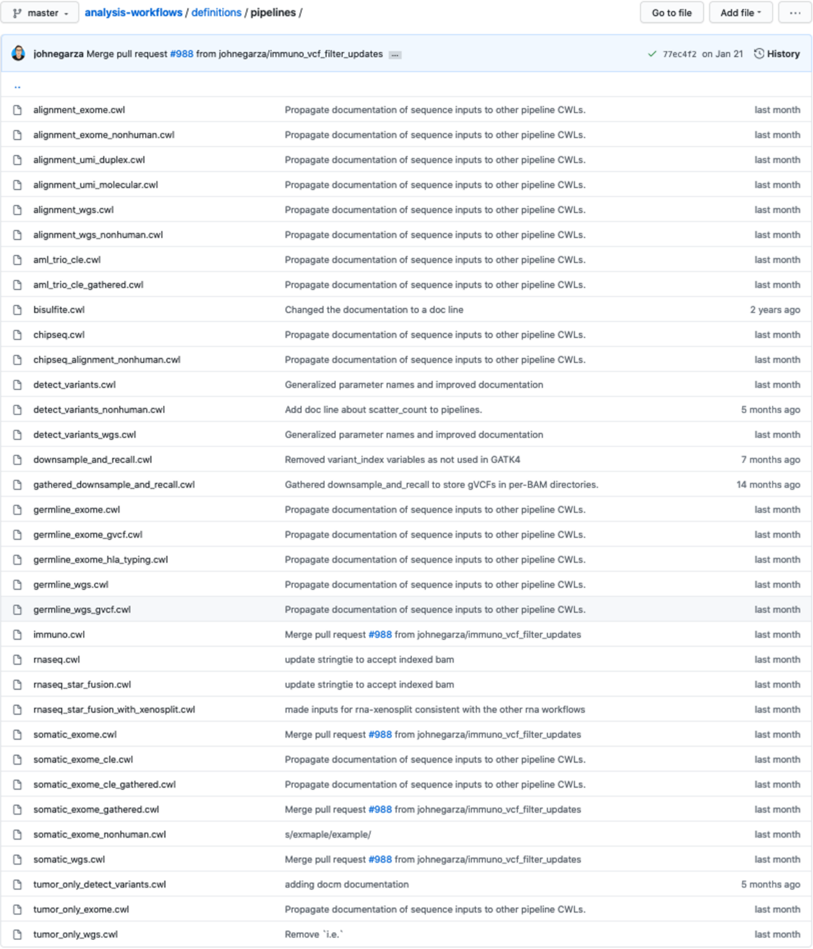
- *Should* be platform-independent

- Workflow systems are complicated

- Our cluster has some unique quirks

- Transition to a new workflow language that fits with our local/GCP model (WDL)

- **griffithlab/analysis-wdls**



analysis-workflows / definitions / pipelines /

johnegarza Merge pull request #988 from johnegarza/immuno_vcf_filter_updates ... ✓ 77ec4f2 on Jan 21

..

alignment_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_exome_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_umi_duplex.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_umi_molecular.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_wgs.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
alignment_wgs_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
aml_trio_cle.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
aml_trio_cle_gathered.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
bisulfite.cwl	Changed the documentation to a doc line	2 years ago
chipseq.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
chipseq_alignment_nonhuman.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
detect_variants.cwl	Generalized parameter names and improved documentation	last month
detect_variants_nonhuman.cwl	Add doc line about scatter_count to pipelines.	5 months ago
detect_variants_wgs.cwl	Generalized parameter names and improved documentation	last month
downsample_and_recall.cwl	Removed variant_index variables as not used in GATK4	7 months ago
gathered_downsample_and_recall.cwl	Gathered downsample_and_recall to store gVCFs in per-BAM directories.	14 months ago
germline_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_exome_gvcf.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_exome_hla_typing.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_wgs.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
germline_wgs_gvcf.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
immuno.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
rnaseq.cwl	update stringtie to accept indexed bam	last month
rnaseq_star_fusion.cwl	update stringtie to accept indexed bam	last month
rnaseq_star_fusion_with_xenosplit.cwl	made inputs for rna-xenosplit consistent with the other rna workflows	last month
somatic_exome.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
somatic_exome_cle.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
somatic_exome_cle_gathered.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
somatic_exome_gathered.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
somatic_exome_nonhuman.cwl	s/exmaple/example/	last month
somatic_wgs.cwl	Merge pull request #988 from johnegarza/immuno_vcf_filter_updates	last month
tumor_only_detect_variants.cwl	adding docm documentation	5 months ago
tumor_only_exome.cwl	Propagate documentation of sequence inputs to other pipeline CWLs.	last month
tumor_only_wgs.cwl	Remove `i.e.`	last month

Workflows on Google Cloud

- Assumes that:
    - you have already set up a PO and billing and have IT support
    - you have access to the cloud console to grant permissions and such

- https://github.com/griffithlab/cloud-workflows