

Malarial Genomics on FireCloud

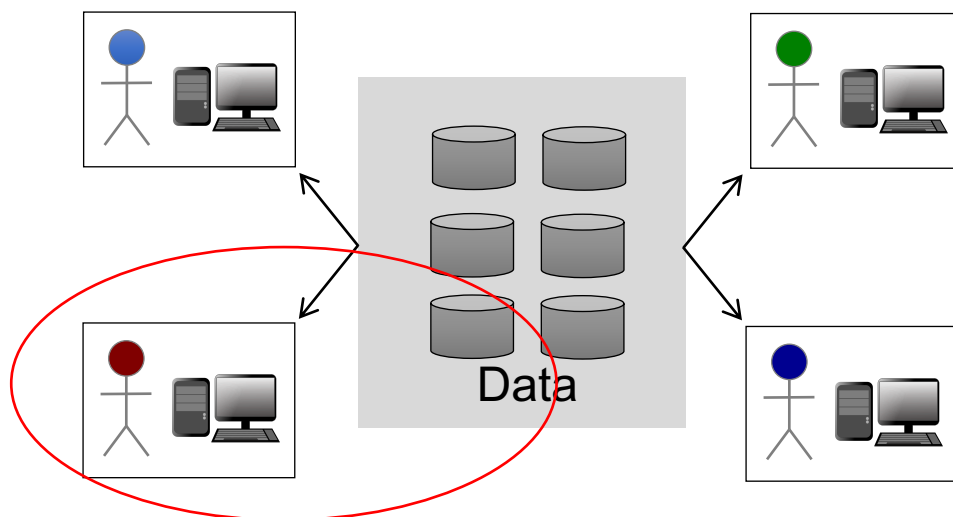
Broad Microbial Group Meeting

Nov 29, 2018

Tim Farrell

Inverting the Model of Genomic Science

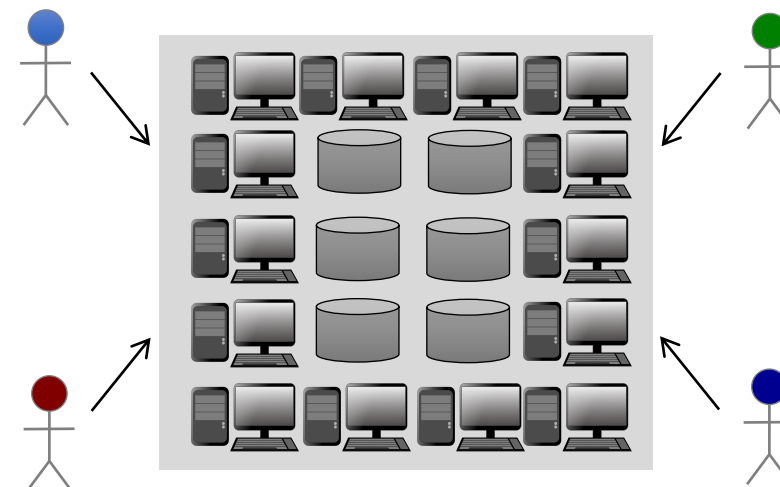
Traditional Approach:
Bring data/ tools to researchers



Problems

- Data Sharing = Data Copying
 - High cost and inconsistency
- Infrastructure Needed
- Siloed Compute
 - Hard to replicate/ reuse

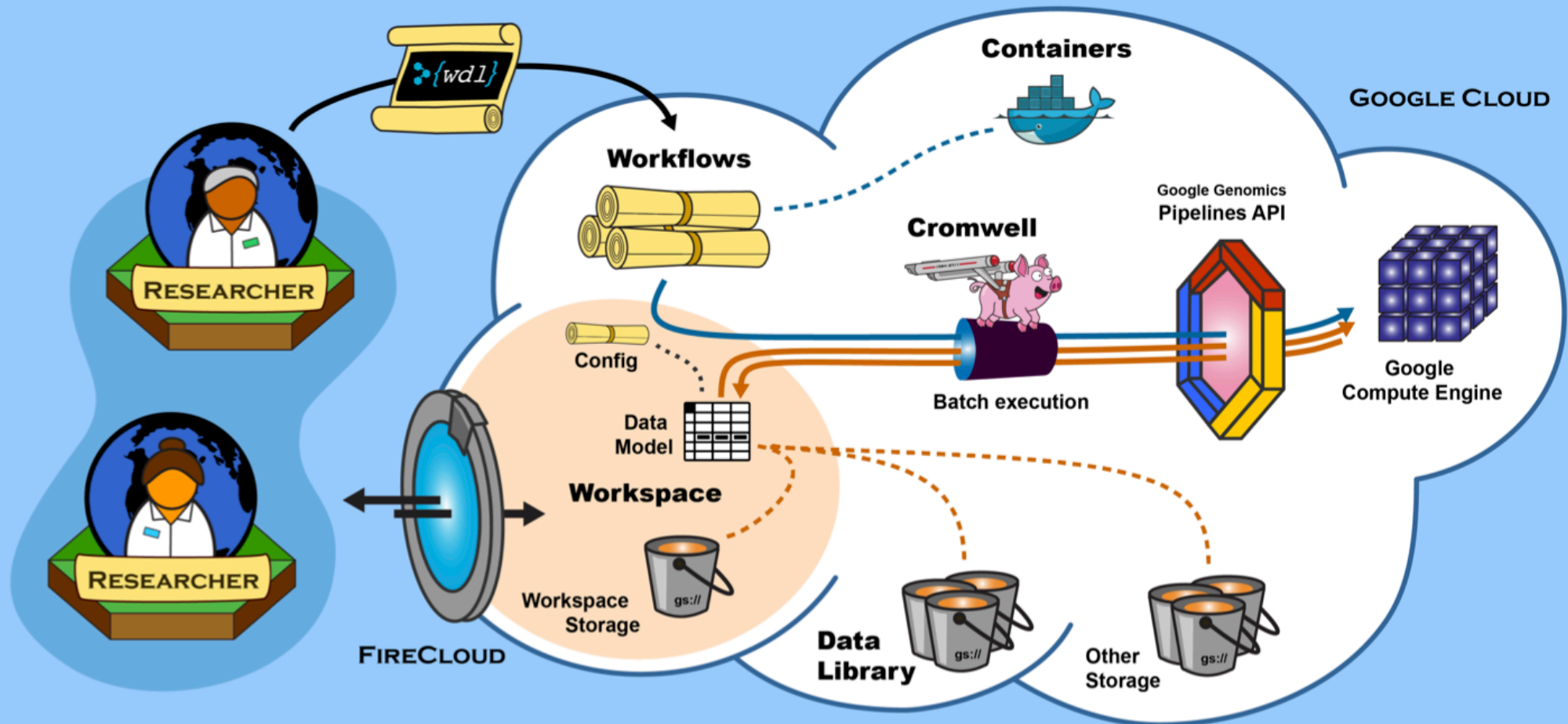
Cloud Approach:
Bring researchers to data/ tools



Advantages

- Cost and Consistency
- Increased Accessibility
 - And more control
- Shared & Elastic Compute

Batch execution of workflows in FireCloud



FireCloud

<https://portal.firecloud.org>

Applicability to Malarial Genomics

Science

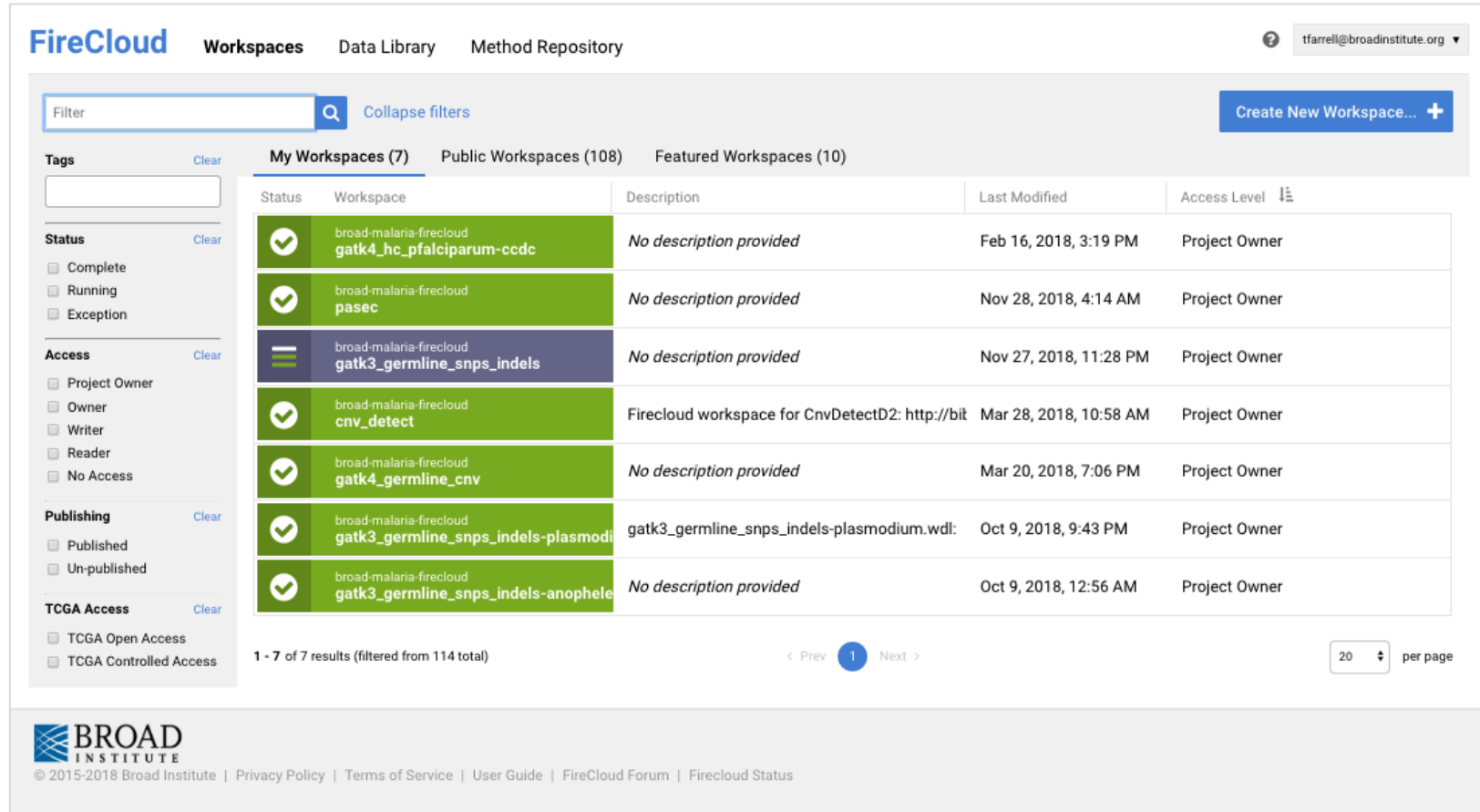
- Data, compute infrastructure and pipelines used both in-house and made available to collaborators
 - Standardized, reusable, transparent workflows
 - Bioinformatics expertise not necessarily required
 - Centralized data store for aggregating datasets

Systems (useful science $\Rightarrow$ automated for routine use)

- Statistics for use by policy makers and clinical decision-makers
- Surveillance
- Prediction/ classification

broad-malaria-firecloud

Workspaces for variant-calling, CNV-calling and amplicon sequencing analysis



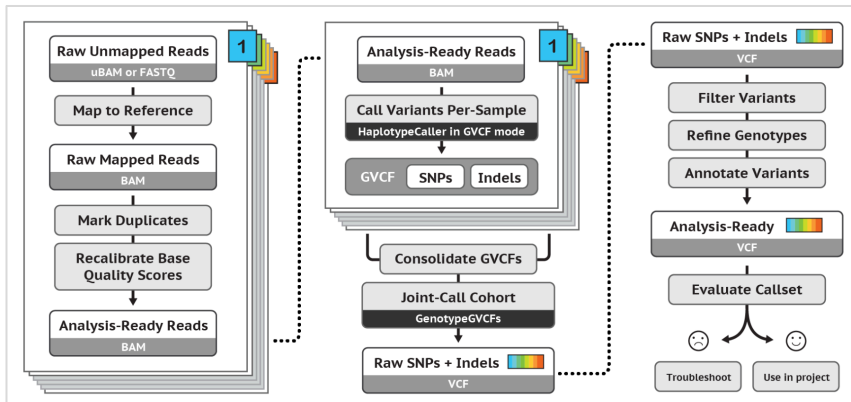
The screenshot displays the FireCloud Workspaces interface. The top navigation bar includes 'FireCloud', 'Workspaces', 'Data Library', and 'Method Repository'. A user profile 'tfarrell@broadinstitute.org' is in the top right. A search bar with 'Filter' and a 'Collapse filters' link is present. A 'Create New Workspace...' button is in the top right corner. The left sidebar contains filters for 'Tags', 'Status' (Complete, Running, Exception), 'Access' (Project Owner, Owner, Writer, Reader, No Access), 'Publishing' (Published, Un-published), and 'TCGA Access' (TCGA Open Access, TCGA Controlled Access). The main content area shows 'My Workspaces (7)', 'Public Workspaces (108)', and 'Featured Workspaces (10)'. A table lists the workspaces:

Status	Workspace	Description	Last Modified	Access Level
✓	broad-malaria-firecloud gatk4_hc_pfalciparum-ccdc	No description provided	Feb 16, 2018, 3:19 PM	Project Owner
✓	broad-malaria-firecloud pasec	No description provided	Nov 28, 2018, 4:14 AM	Project Owner
≡	broad-malaria-firecloud gatk3_germline_snps_indels	No description provided	Nov 27, 2018, 11:28 PM	Project Owner
✓	broad-malaria-firecloud cnv_detect	Firecloud workspace for CnvDetectD2: http://bit	Mar 28, 2018, 10:58 AM	Project Owner
✓	broad-malaria-firecloud gatk4_germline_cnv	No description provided	Mar 20, 2018, 7:06 PM	Project Owner
✓	broad-malaria-firecloud gatk3_germline_snps_indels-plasmodi	gatk3_germline_snps_indels-plasmodium.wdl:	Oct 9, 2018, 9:43 PM	Project Owner
✓	broad-malaria-firecloud gatk3_germline_snps_indels-anophele	No description provided	Oct 9, 2018, 12:56 AM	Project Owner

At the bottom, it shows '1 - 7 of 7 results (filtered from 114 total)' and a pagination control set to '20 per page'. The Broad Institute logo and footer information are at the very bottom.

broad-malaria-firecloud/gatk3_germline_snps_indels

- Configurations for parasite and vector variant-calling
- Parallelizes over samples as well as over genomic intervals



The screenshot shows the FireCloud interface for the workspace 'broad-malaria-firecloud/gatk3_germline_snps_indels'. The 'Method Configurations' tab is active, displaying a table of configurations.

Name	Root Entity Type	Method Source	Method
gatk3_germline_snps_indels-anopheles	sample_set	FireCloud	broad-malaria-firecloud-methods/gatk3_germline_snps_indels Snapshot ID: 8
gatk3_germline_snps_indels-pfalciparum	sample_set	FireCloud	broad-malaria-firecloud-methods/gatk3_germline_snps_indels Snapshot ID: 8
gatk3_germline_snps_indels-pviva	sample_set	FireCloud	broad-malaria-firecloud-methods/gatk3_germline_snps_indels Snapshot ID: 8

Below the table, it shows '1 - 3 of 3 results' and a pagination control set to '20 per page'. The footer includes the Broad Institute logo and copyright information: '© 2015-2018 Broad Institute | Privacy Policy | Terms of Service | User Guide | FireCloud Forum | Firecloud Status'.

broad-malaria-firecloud/gatk3_germline_snps_indels

- 96 anopheles bams, avg. 10 GB
- \$400 for compute (\$4/ bam)
- 30 hours to compute (10 hrs of variant QC, since been parallelized)

WORKSPACE
broad-malaria-firecloud/gatk3_germline_snps_indels-anopheles

Summary Data Analysis Notebooks **BETA** Method Configurations Monitor

 Done

Method Configuration

Namespace: broad-malaria-firecloud-methods
Name: gatk3_germline_snps_indels-anopheles-pignatelli-1_SMv30rH65aU ⓘ

Submitted by

tfarrell@broadinstitute.org
September 21, 2018, 2:09 PM (2 months ago)

Submission ID

[e0eb8400-92e7-4475-8c69-438f06ee823e](#) ⓘ

Total Run Cost ⓘ

\$385.33

Call Caching ⓘ

Disabled

Submission Entity

Type: sample_set
Name: full

Workflows:

Workflows > full

Workflow ID: [412c262c-d784-481f-bac5-7f067f67b03b](#) ⓘ

Status:  Succeeded

Total Run Cost: \$385.33

Submitted: September 21, 2018, 2:11 PM (2 months ago)

Started: September 21, 2018, 2:36 PM (2 months ago)

Ended: September 22, 2018, 8:03 PM (2 months ago)

Inputs: [Show](#)

Outputs: [Show](#)

Workflow Log: [workflow.412c262c-d784-481f-bac5-7f067f67b03b.log](#)

Workflow Timing: [Show](#)

Calls:

-  [GATK3_Germline_Variants.HaplotypeCallerGvcf_GATK3](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.GenotypeGVCFs](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.HardFiltration](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.CombineGVCFs](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.MarkDuplicates](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.ReorderBam](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.GatherVCFs](#) ⓘ [Show](#)
-  [GATK3_Germline_Variants.SnpEff](#) ⓘ [Show](#)