GATK Best Practices for Variant Discovery

GATK4 Workshop

Montreal, Quebec
20-23 March, 2018

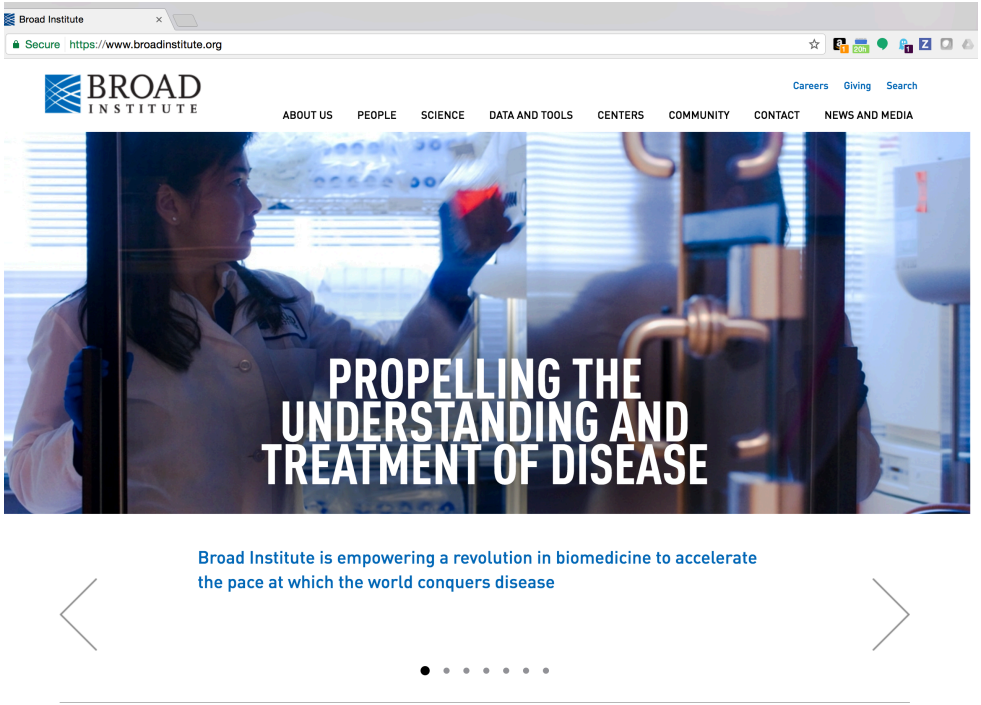
Data Sciences Platform
Broad Institute of Harvard and MIT
https://software.broadinstitute.org/gatk/

BROAD INSTITUTE

@gatk_dev
@wdl_dev

gatk

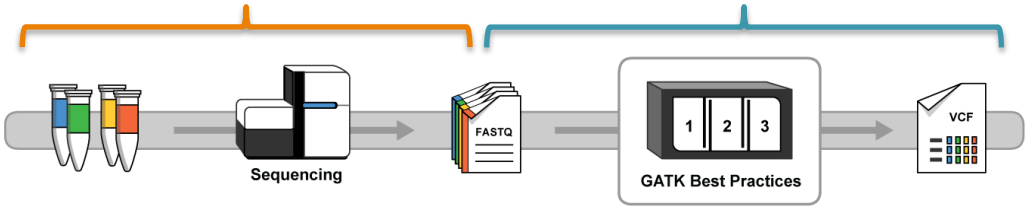
What / who is the Broad Institute?



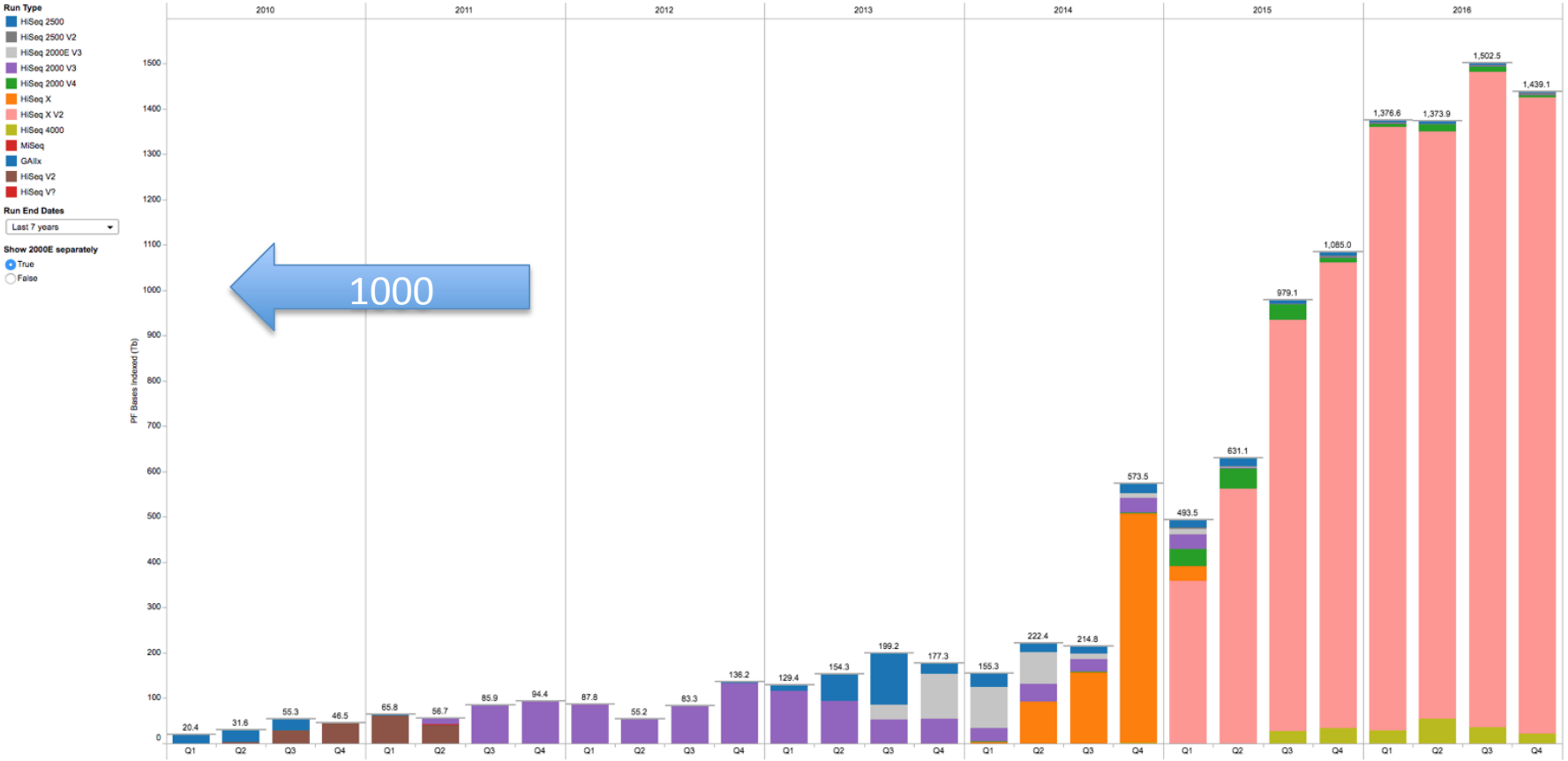
- Non-profit research institution
- Spinoff of Harvard & MIT
- Eric Lander and philanthropists Eli & Edyth Broad (and others)
- Aims to use the full power of genomics to transform the understanding and treatment of disease

We are Broad Genomics

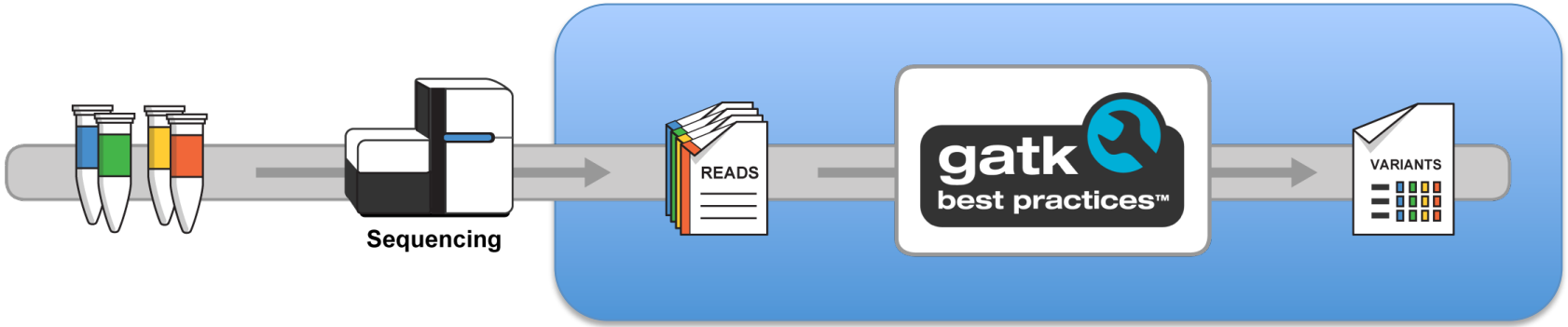
Genomics Platform + Data Sciences Platform

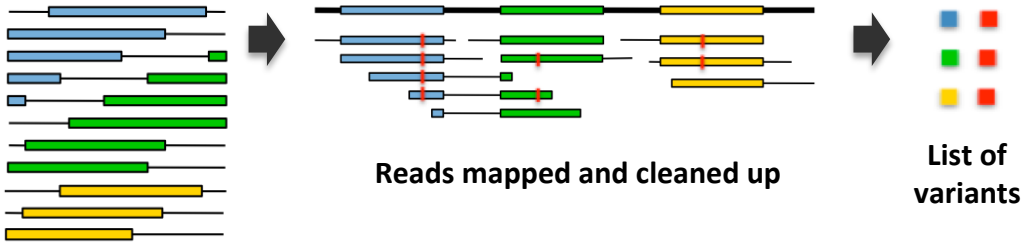


Quarterly output (in TBases) of the Genomics Platform



Variant discovery with GATK





Enormous pile of short reads

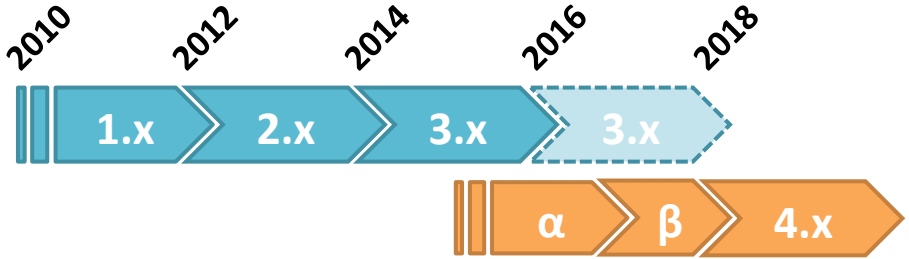
Reads mapped and cleaned up

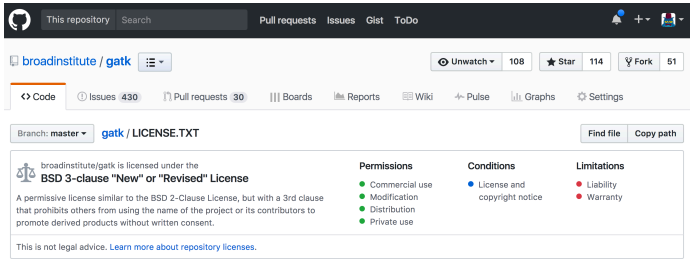
List of variants

Introducing GATK version 4.0

- Re-engineered for **speed**, **scalability** and **versatility**
- Expanded **scope of analysis** to more variant types
- Reproducible **best practices workflows**

gatk best practices™

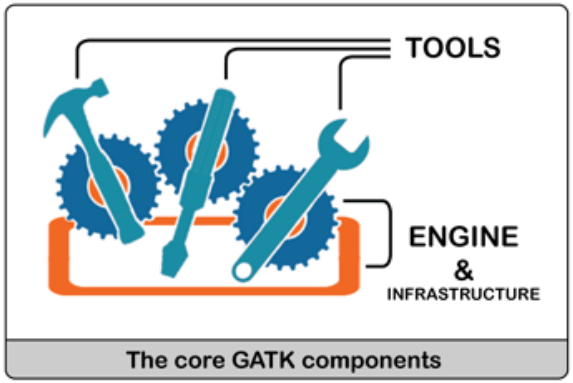




Full open-source under BSD 3-clause

New engine rewritten from scratch

Completely re-implemented to enable dramatic performance improvements plus major new functionality and analytical capabilities

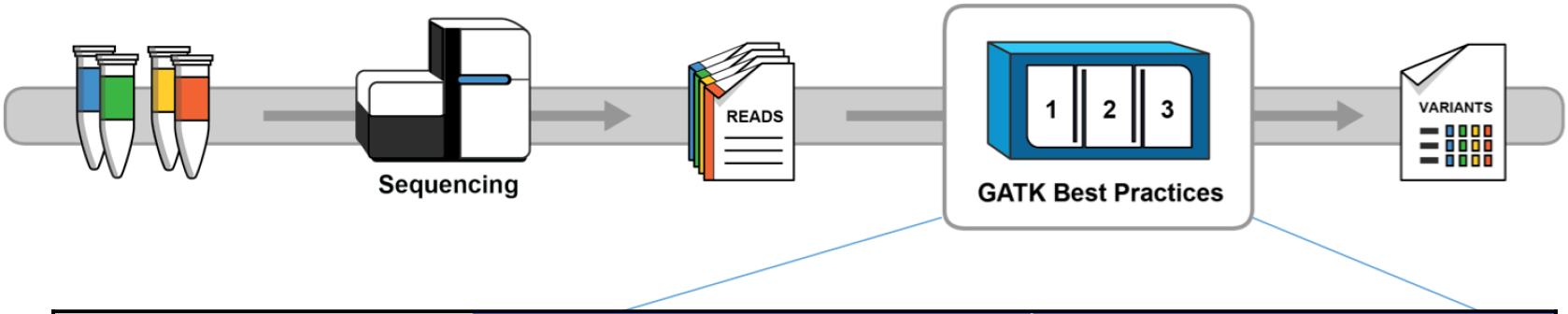


The core GATK components

- Streamlined architecture ➔ overall efficiency
- Intel Genomics Kernel Library (GKL) ➔ speed
- Intel GenomicsDB ➔ scalability
- Apache Spark support ➔ robust parallelism
- Google Dataproc and GCS support ➔ cloud execution
- Versatility of data traversal ➔ analysis scope

With invaluable help from intel cloudera Google Cloud Platform

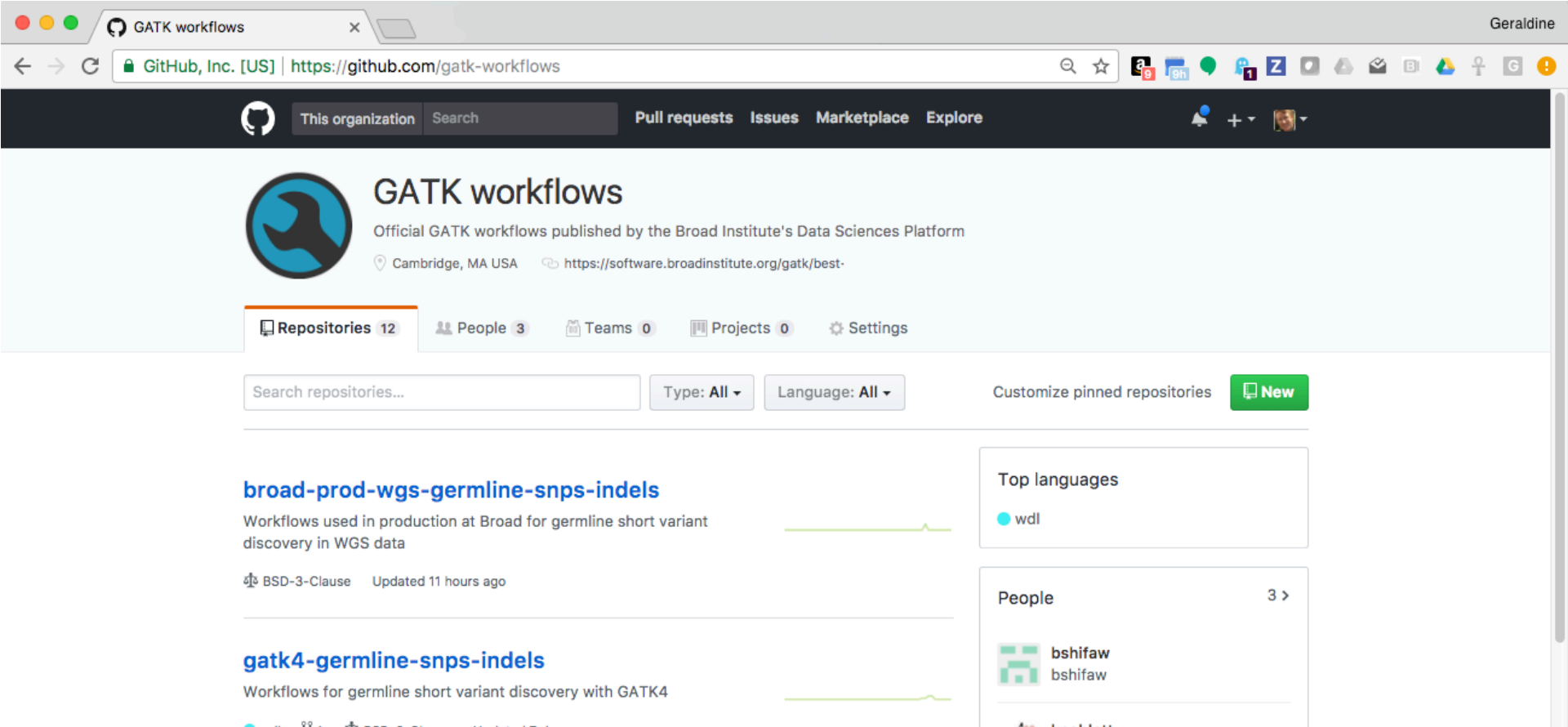
Best Practices workflows for more use cases



	GERMLINE	SOMATIC
SNPs & INDELs	**HaplotypeCaller GVCF**	**MuTect2**
Copy Number	**GATK gCNV** (beta)	**GATK CNV + aCNV**
Structure Variation	**GATK SVDiscovery** (beta)	(planned)

Established tools / New tools

Workflow scripts (WDL) deposited in GitHub



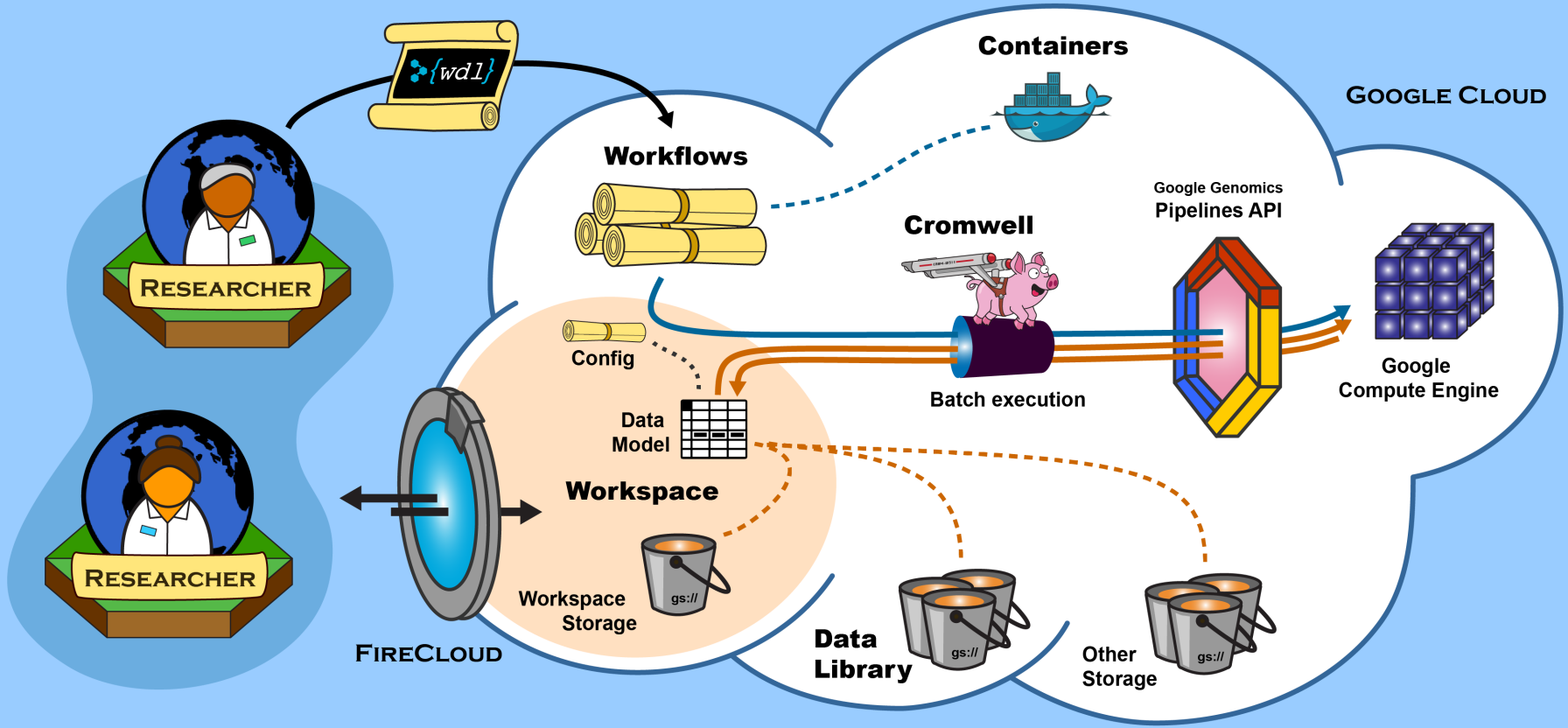
WDL: a workflow language that humans can understand

```
workflow myWorkflowName {
    File my_ref
    File my_input
    String name

    call task_A {
        input: ref= my_ref, in= my_input, id= name
    }
    call task_B {
        input: ref= my_ref, in= task_A.out
    }
}
task task_A { ... }
task task_B { ... }
```

```
task task_A {
    File ref
    File in
    String id

    command {
        do_stuff -R ${ref} -I ${in} -O ${id}.ext
    }
    runtime {
        docker: "my_project/do_stuff:1.2.0"
    }
    output {
        File out= "${id}.ext"
    }
}
```

http://www.openwdl.org

FireCloud puts GATK4 workflows in everyone's hands



Workshop schedule

	Morning	Afternoon
Day 1	Introductions	Pre-Processing
Day 2	Germline Variant calling	Germline Variant Filtering Callset Evaluation
Day 3	Somatic small variants (SNVs and Indels)	Somatic large variants (CNVs and SVs)
Day 4	Pipelining with Cromwell	Working with Firecloud