



# GATK Best Practices for Variant Discovery



## GATK4 Workshop

Taipei City, Taiwan  
6-9 March, 2018

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



@gatk\_dev  
@wdl\_dev

iSEQTOOLS  
<http://iseqtools.org/>

# What / who is the Broad Institute?



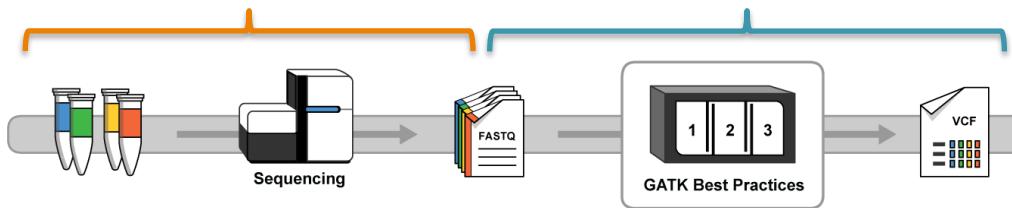
The screenshot shows the homepage of the Broad Institute website at https://www.broadinstitute.org. The page features a large banner image of a scientist in a lab coat working in a laboratory. Overlaid on the banner is the text "PROPELLING THE UNDERSTANDING AND TREATMENT OF DISEASE". Below the banner, a subtitle reads "Broad Institute is empowering a revolution in biomedicine to accelerate the pace at which the world conquers disease". The website navigation bar includes links for ABOUT US, PEOPLE, SCIENCE, DATA AND TOOLS, CENTERS, COMMUNITY, CONTACT, and NEWS AND MEDIA. The top right of the page includes links for Careers, Giving, and Search, along with social media icons.

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

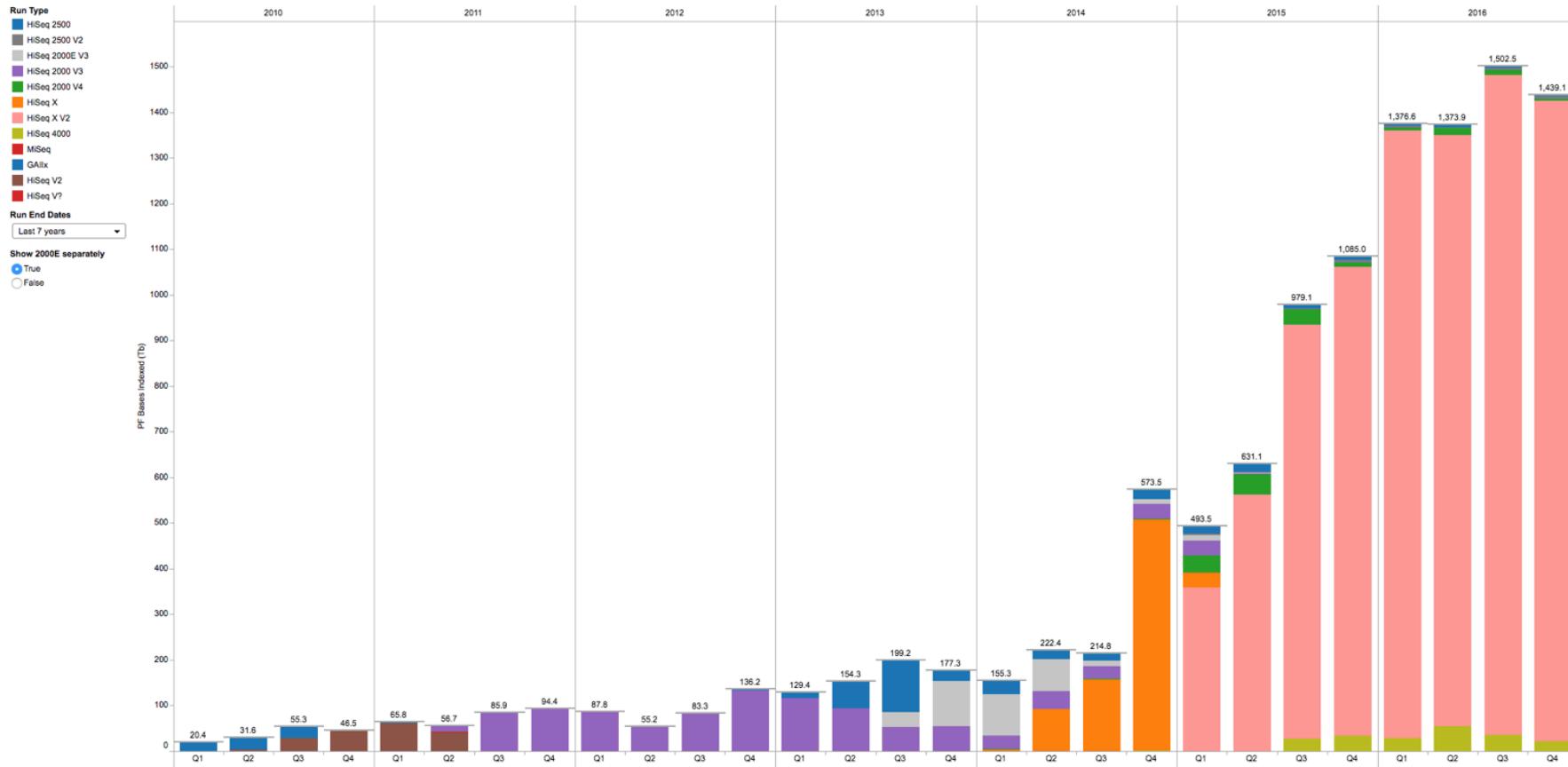
# We are Broad Genomics



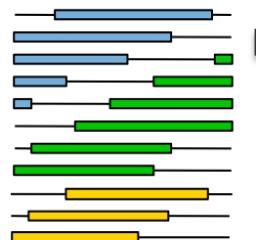
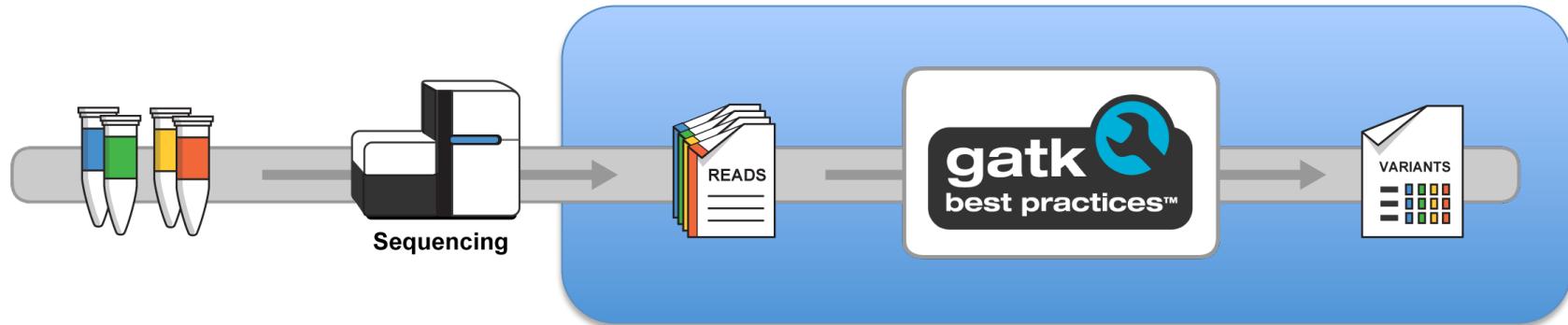
Genomics Platform + Data Sciences Platform



# Quarterly output (in TBases) of the Genomics Platform



# Variant discovery with GATK



Enormous pile  
of short reads

Reads mapped and cleaned up

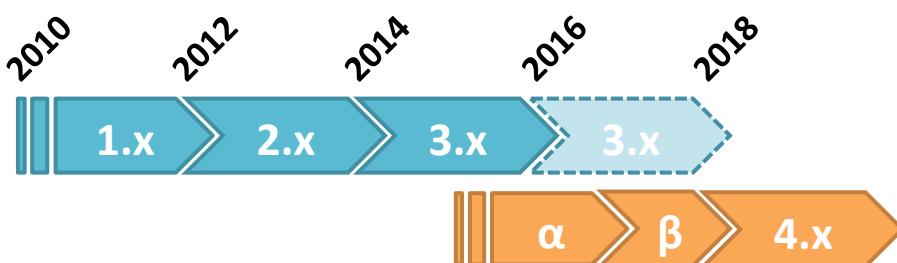


List of  
variants

# Introducing GATK version Four Point Ohhh



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



This repository Search Pull requests Issues Gist To Do

broadinstitute / gatk

Code Issues 430 Pull requests 30 Boards Reports Wiki Pulse Graphs Settings

Branch: master gatk / LICENSE.TXT

Find file Copy path

broadinstitute/gatk is licensed under the **BSD 3-clause "New" or "Revised" License**

A permissive license similar to the BSD 2-Clause License, but with a 3rd clause that prohibits others from using the name of the project or its contributors to promote derived products without written consent.

Permissions

- Commercial use
- Modification
- Distribution
- Private use

Conditions

- License and copyright notice

Limitations

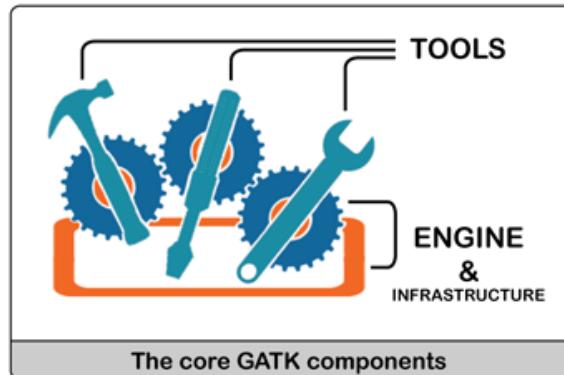
- Liability
- Warranty

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**



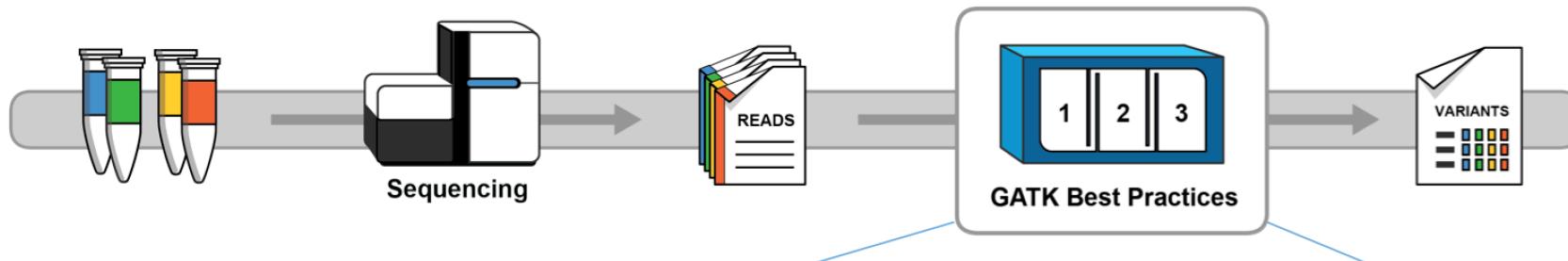
- Streamlined architecture → overall efficiency
- Intel Genomics Kernel Library (GKL) → speed
- Intel GenomicsDB → scalability
- Apache Spark support → robust parallelism
- Google Dataproc and GCS support → cloud execution
- Versatility of data traversal → analysis scope

*With invaluable help from*



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

The screenshot shows a web browser window with the title bar "GATK workflows". The address bar displays "GitHub, Inc. [US] | https://github.com/gatk-workflows". The page content is the GitHub organization profile for "GATK workflows".

**GATK workflows**  
Official GATK workflows published by the Broad Institute's Data Sciences Platform  
Cambridge, MA USA | https://software.broadinstitute.org/gatk/best-

**Repositories** 12 | **People** 3 | **Teams** 0 | **Projects** 0 | **Settings**

Search repositories... | Type: All | Language: All | Customize pinned repositories | **New**

**broad-prod-wgs-germline-snps-indels**  
Workflows used in production at Broad for germline short variant discovery in WGS data  
BSD-3-Clause | Updated 11 hours ago

**Top languages**  
wdl

**gatk4-germline-snps-indels**  
Workflows for germline short variant discovery with GATK4

**People** 3 >  
bshifaw bshifaw

# 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"
```

```
        }
```

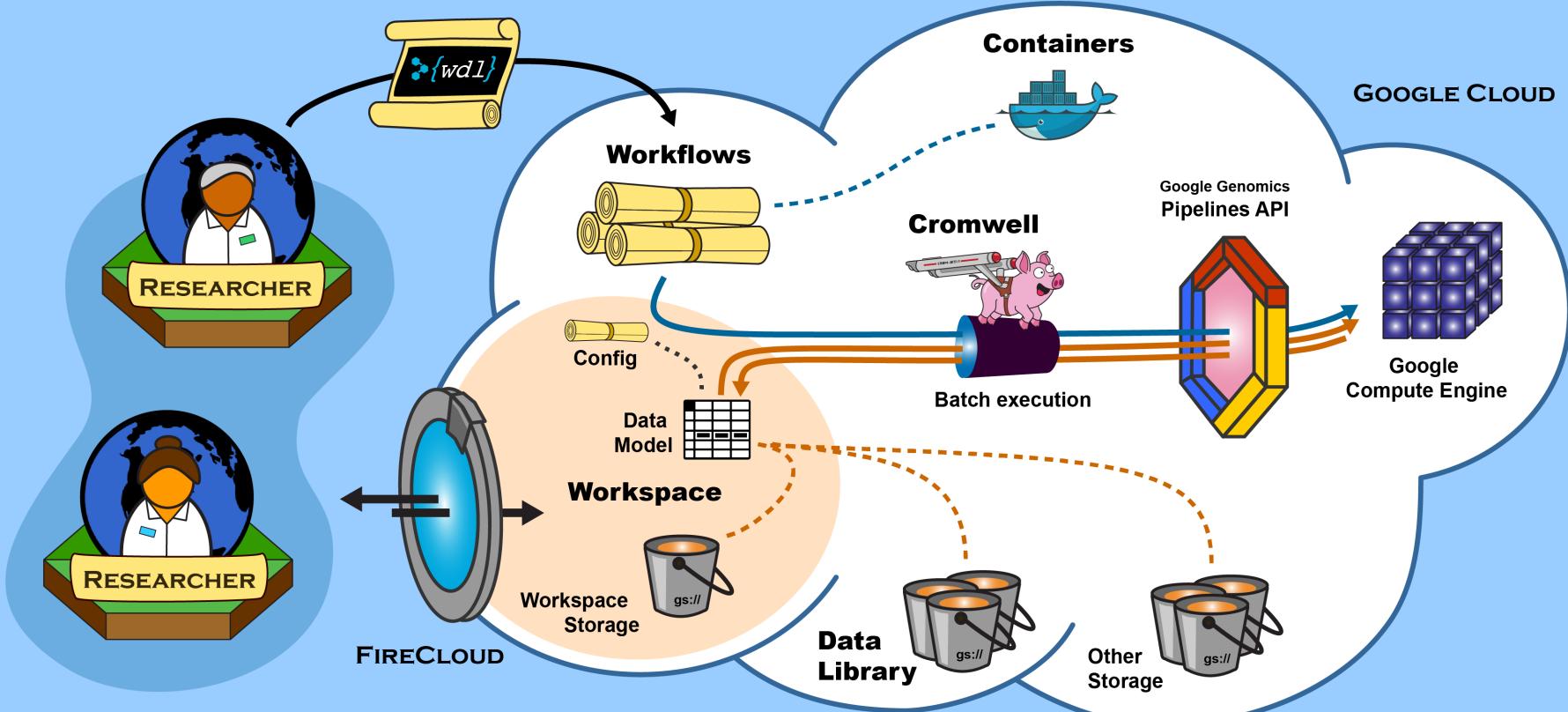
```
        output {
```

```
            File out= "${id}.ext"
```

```
        }
```

```
}
```

# FireCloud puts GATK4 workflows in everyone's hands



# Workshop schedule



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