GATK & Terra

A brief history
Enormous pile of short reads

Reads mapped and cleaned up

List of variants

GATK = Genome Analysis Toolkit
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>GATK 1.0 framework</td>
<td>Open source (MIT) 1000 Genomes Project</td>
</tr>
<tr>
<td>2010</td>
<td>GATK 2.0 toolkit rebranding</td>
<td>NHGRI iSeqTools</td>
</tr>
<tr>
<td>2011</td>
<td>GATK 3.0 scaling up</td>
<td>ExAC + gnomAD</td>
</tr>
<tr>
<td>2012</td>
<td>GATK 4 BETA re-engineering</td>
<td>Hybrid licensing with partner (Appistry) Hybrid licensing @Broad (all in-house)</td>
</tr>
<tr>
<td>2013</td>
<td>GATK 4.0 cloud support &amp; scope expansion</td>
<td>Open source (BSD-3)</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td>Intel-Broad Center</td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td>Google Cloud Partnership</td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td>Cloud 6 Partnership</td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td>Alibaba Cloud</td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td>ILMN DRAGEN</td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The real story is in the cloud
Explosive growth in data generation
Pipeline throughput is "spiky"

Genome processing requests per day over several months

- **Advantages of cloud over on-premises computing:**
  - No need to pay for compute power when we aren’t using it
  - Can tolerate spikes without being forced to maintain a backlog of "things to process once everything calms down"

  + No longer need to take over Allston, MA with giant datacenter
Data access and federation are key

### Traditional Way
Bring data to the researchers

- Data Repository

### Problems
- Data sharing = data copying
- Requires big infrastructure at each site
- Largely fixed compute
- Individual security implementations

### Cloud Way
Bring researchers to the data

- Public Cloud

### Solutions
- True data sharing
- Cloud provides the infrastructure
- Elastic compute and storage
- Centralized security implementation
Data Sciences Platform developed FireCloud

- Collaborative **cloud-based** analysis platform built on top of **Google Cloud Platform**
- **Free to access** / compute & storage charged by Google
- Access published **data** and/or add your own
- Access existing **methods** and/or add your own
- **Execute** analyses in auditable pipelines
- **Share** data, methods and results with collaborators
Next step: Generalize platform model for data access & analysis
Welcome to Terra

Terra is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate.

Find how-to’s, documentation, video tutorials, and discussion forums

Learn more about the Terra platform and our co-branded sites

View Workspaces
Workspace connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.

View Examples
Browse our gallery of showcase Workspaces to see how science gets done.

Browse Data
Access data from a rich ecosystem of data portals.

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 1740023 under Contract No. HHSN261200800001E
Terra supports the full lifecycle of data access & analysis
Terra connects to datasets hosted by many orgs/apps

- NCI Cloud Pilot (TCGA)
- Nurses Health Study
- Human Cell Atlas
- NIH Data Commons (TopMed, GTEx)
- NHGRI AnVIL (CCGD, CMG, eMERGE)
Data Biosphere
A shared vision for an open ecosystem

• Modular
• Open
• Community-driven
• Standards-based

A Data Biosphere for Biomedical Research
We, the authors listed below, are privileged to be part of the growing global community bringing data and life science together. Our groups have been

Read the manifesto on MEDIUM
broad.io/databiosphere
That’s the story so far...