**Geraldine A Van der Auwera**

**Broad Institute**

**GATK and Terra**

**1/8/2020**

GATK: Genome Analysis Toolkit

Main focus on variant discovery

GATK 1.0 2009 – 2011, framework, MIT License

GATK 2.0 2012 – 2014, toolkit rebranding, Hybrid licensing with partner (Appistry), NHGRI iSeqTools. Free for academic researchers, commercial license for for-profit users

GATK 3.0 2014 – 2015, scaling up, fueled by ExAC + gnomAD

2016 – 2017 GATK 4 BETA, re-engineering, hybrid licensing @ Broad (all in-house)

GATK 4.0 – 2018 and beyond, cloud support and scope expansion, Open source (BSD-3)

Partnerships:

Intel-Broad Center – Contract that funds development of GATK

Google Cloud Partnership – No direct funding, but GCP provided engineering help to move Broad pipelines to the cloud

Cloud 6 Partnership – GCP, MSFT Azure, AWS, IBM, Cloudera, Intel

Alibaba Cloud – Main cloud provider in China

ILMN DRAGEN partnership

FireCloud – 2016

Terra – 2019

Broad pipeline throughput is “spiky” so advantages of cloud over on-premises computing

* No need to pay for compute power when not using it
* Data access and federation are key

Cloud way

* True data sharing
* Cloud provides the infrastructure
* Elastic compute and storage
* Centralized security implementation

CD: What did the partnerships provide?

GVA: Google willing to engage in a deep partnership. Sitting together for co-development and training. Helped optimize pipelines. E.g. first pipeline ports cost $45/genome. After optimization, $5/genome.

CD: Does this approach scale?

GVA: Key is to generalize the take-aways and provide guides that other development teams can follow.

CD: Does this scale across tool types?

GVA: Hard to say, but there may be a few different profiles.

CD: Is there benefit to involving industrial partners earlier in development? Would that avoid some of the re-engineering?

GVA: Having an incubation team focused on the scientific questions is useful. Need a phase of experimenting and getting the science right, without caring about the scaling problems. Need to identify when to pivot and productionize the software including usability, portability, reproducibility, scalability. May be a different skillset and can benefit from an industry partnership. Broad benefitted from forming a proper software shop. Could not have been successful at GATK 4 without Intel and Cloudera input.

CD: What was in it for Appistry?

GVA: Partnership didn’t work very well – not enough value on either side. They didn’t provide much on the engineering side, just repackaging. The sales incentive was not appropriate – sometimes got sold for the wrong use case. In the end, no value to Broad or user community.

GS: More about problems with hybrid licensing model?

GVA: Non-profit and for-profit was hard to distinguish. E.g., what about a data processing center within a hospital? Small start-ups, having the extra cost was difficult financially. Couldn’t drop the cost enough to make it affordable for them. Every license was executing a contract, which was a lot of person time.

CD: Can you speak to FDA clearance?

GVA: The free pipelines are “for research use only”. In house, have a CLIA lab that has gone through the process to get certifications for certain versions of the pipelines. These are not distributed. There are companies that are repackaging for clinical use. No instance of this causing restrictions.

MR: How did you initiate some of these partnerships?

GVA: Understanding that a lot of it was conversations Anthony Philipakis had via his networking contacts. Made the business case of what was in it for them. Might be hard for a PI at another institution to pull off. Might need an incubator to support this.

GVA: Funding is from Intel; Broad’s funding from their genomic sequencing service; scientific projects; HCA; others.

<back to slides>

FireCloud was funded through the NCI Cloud Pilots program. Focused on cancer genomics

Next step: Generalize the platform model for data access and analysis.

Data Generators > TERRA < Tool developers

Data Sciences Platform team at Broad develops Terra with Verily. They provide funding and engineering.

Terra is sustainable because it connects to datasets hosted by many orgs/apps. Part of a larger Data Biosphere.

Verily uses the Terra platform to support their customers, so mutual interest.