



Genomics in the Cloud

Book Club - Week 2

December 7, 2020

Agenda

- Chapter 1: Introduction
- Additional resources
- Anything else

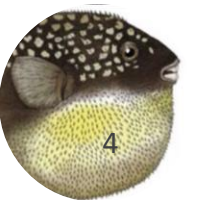


Chapter 1: Introduction

Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

GITC overview

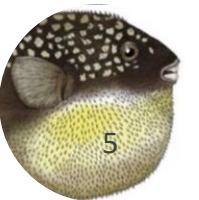
- Essential genomics and computing technology background
- Basic cloud computing operations
- Getting started with GATK
- Three major GATK Best Practices pipelines for variant discovery
- Automating analysis with scripted workflows using WDL and Cromwell
- Scaling up workflow execution in the cloud, including parallelization and cost optimization
- Interactive analysis in the cloud using Jupyter notebooks
- Secure collaboration and computational reproducibility using Terra



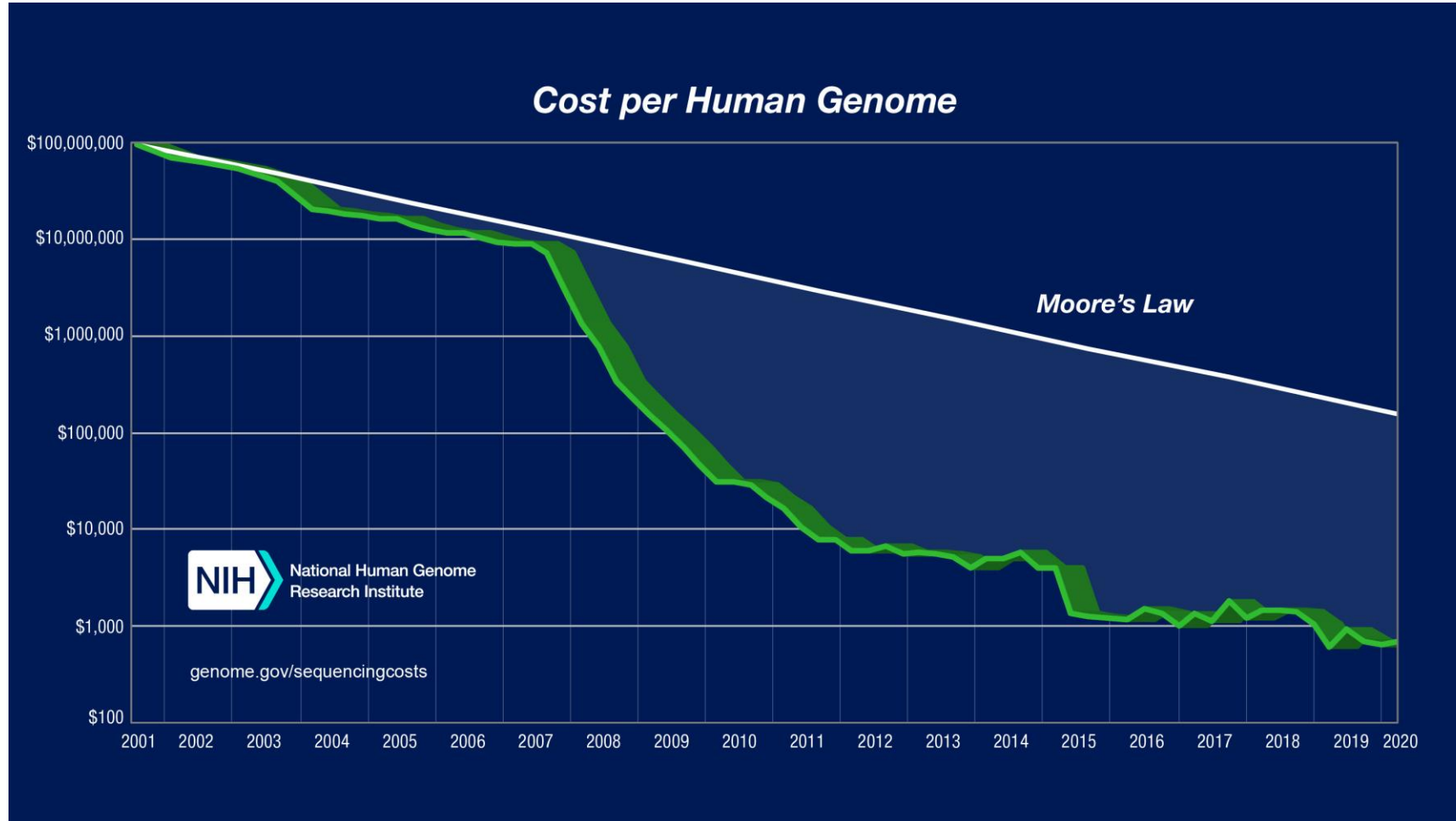
Tools and methods



- Genome Analysis Toolkit (GATK)
 - <https://gatk.broadinstitute.org/hc/en-us>
- Workflow Description Language (WDL)
 - <https://openwdl.org>
- Google Cloud Platform (GCP)
 - <https://cloud.google.com>
- Terra
 - <https://terra.bio>



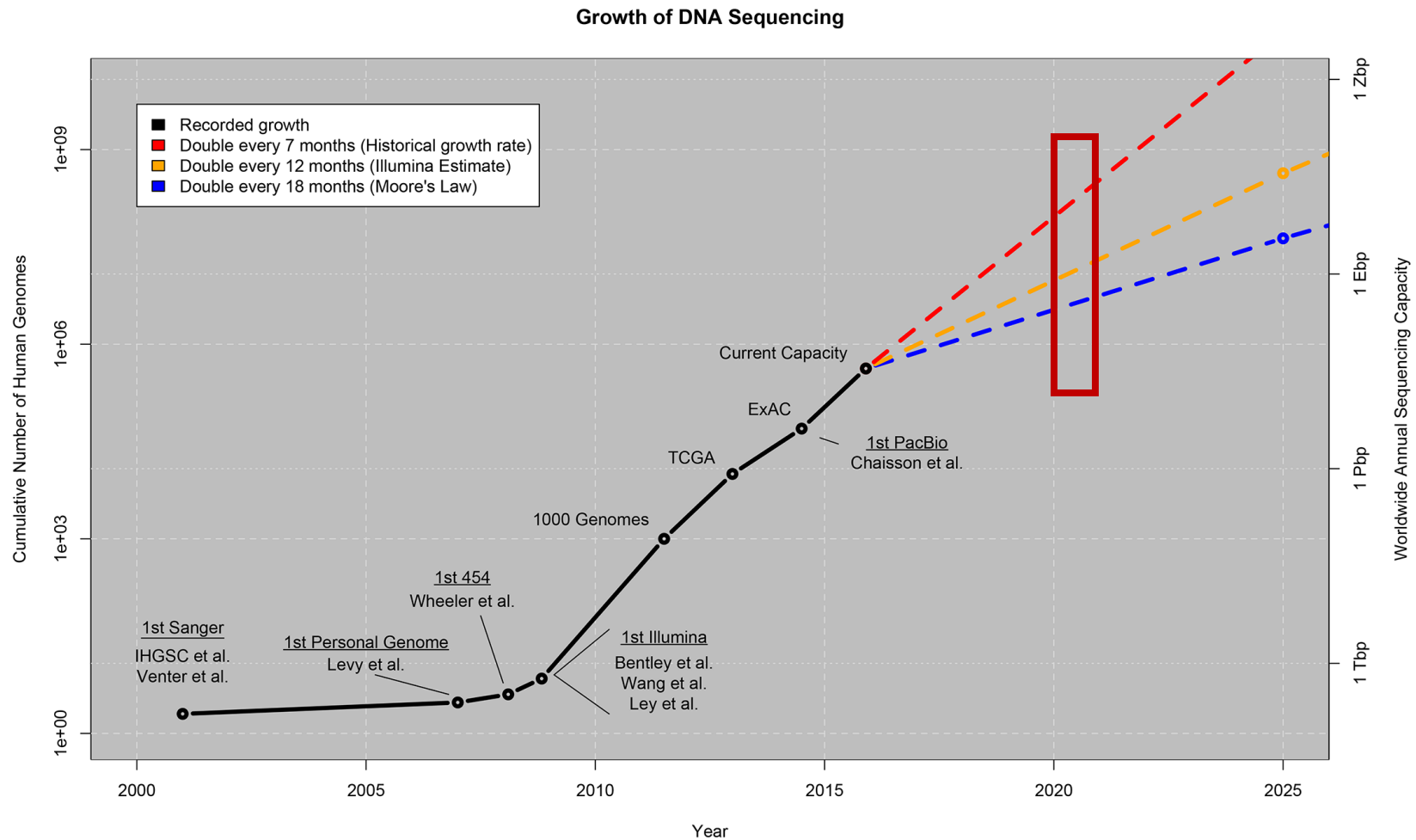
Sequencing costs: from \$3B to <\$1,000 in 20 years



<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>



An explosion of data



Billions and
billions of
genomes

<https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195>



Why cloud?

Traditional approach

Bring data to researchers

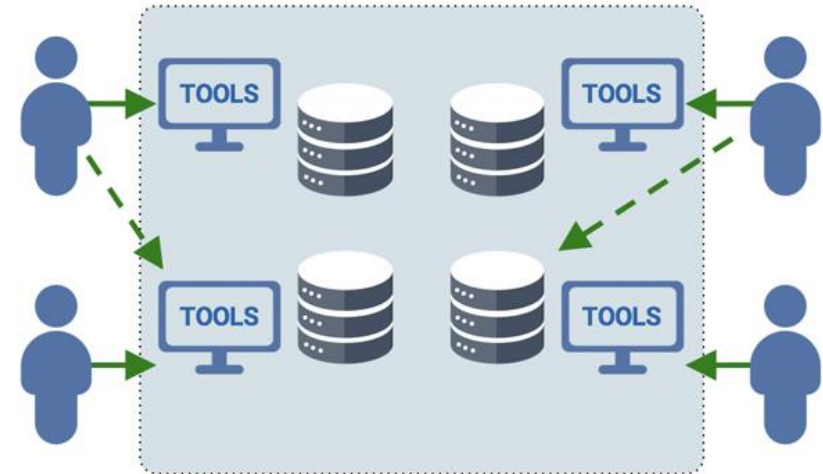


Discourages shared research

- "Weakest link" security
- Huge infrastructure needed
- Pay for multiple copies
- Bespoke & unsupported tools

Cloud-centric approach

Bring researchers to data

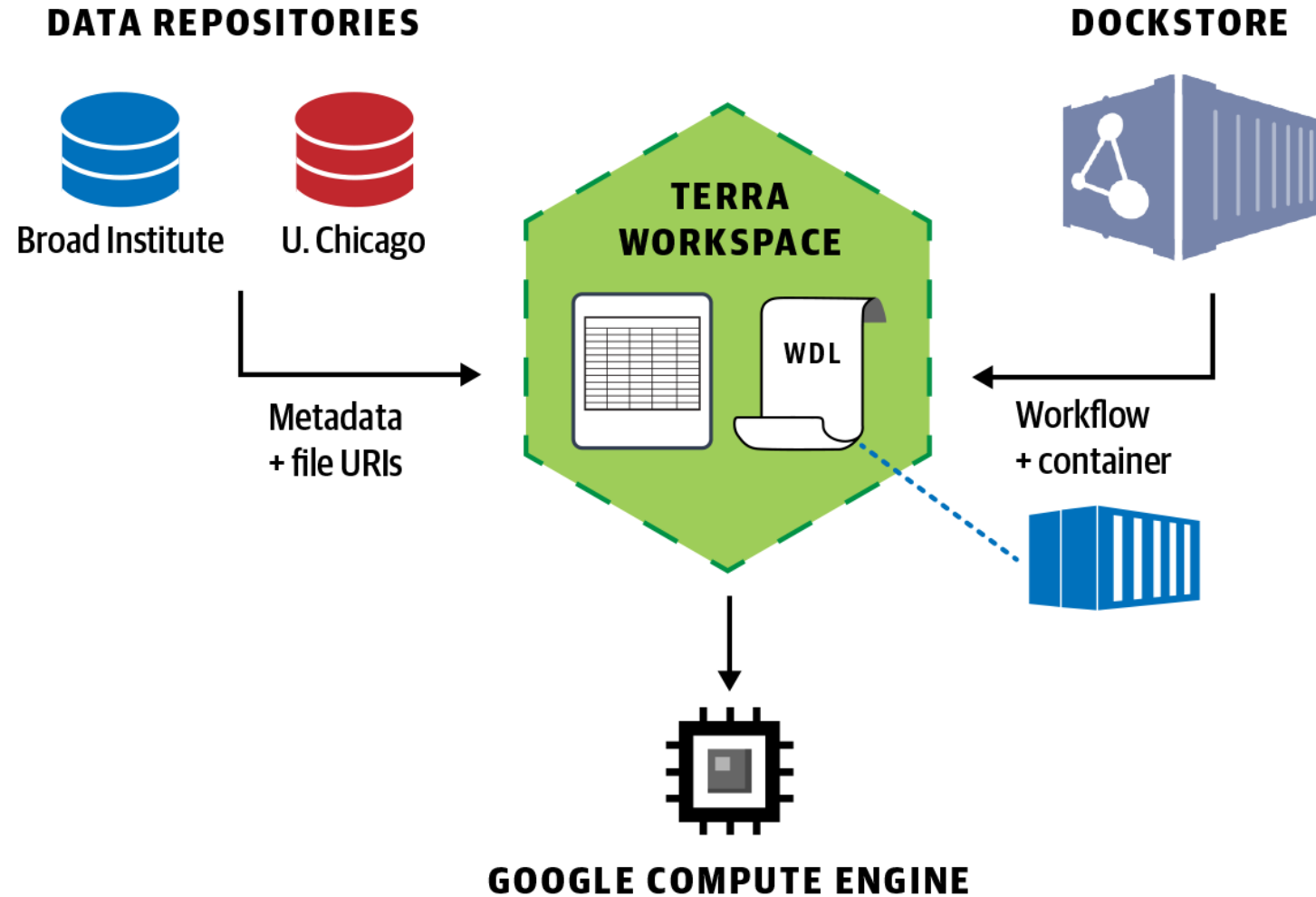


Facilitates collaboration

- Centralized security controls
- Accessible to all researchers
- Decreased cost of storage
- Shared tool ecosystem

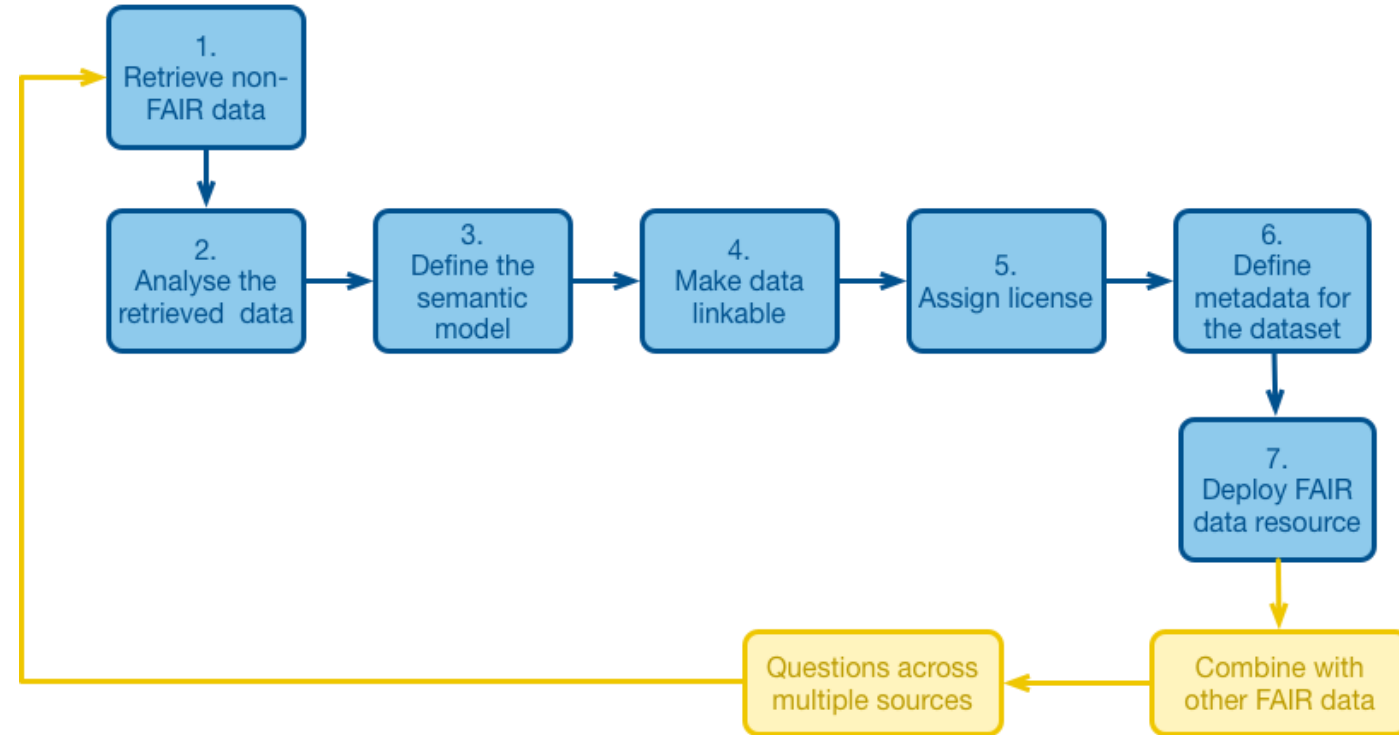


Reusable infrastructure



Making it FAIR

- Findable
- Accessible
- Interoperable
- Reusable



Additional resources

- Source code and related materials
 - <https://github.com/broadinstitute/genomics-in-the-cloud>
- Blog
 - <https://broadinstitute.github.io/genomics-in-the-cloud/>
- Global Alliance for Genomics and Health
 - <https://www.ga4gh.org>
- Chapters / presenters
 - See Slack for link



Chapter	Title	Presenter	Meeting	Date
0	Welcome	KT	1	2020-11-30
1	Intro	KT	2	2020-12-07
2	Genomics		3	2020-12-14
3	Computing		4	2020-12-21
4	Cloud intro		5	2020-12-28
5	GATK intro	Mikhael	6	2021-01-04
6	GATK germline		7	2021-01-11
7	GATK somatic		8	2021-01-18
8	WDL		9	2021-01-25
9	WDL examples		10	2021-02-01
10	Cromwell		11	2021-02-08
11	Terra		12	2021-02-15
12	Jupyter	Joris	13	2021-02-22
13	Terra workspace		14	2021-03-01
14	Reproducibility	Matthieu	15	2021-03-08





Thank you for joining us today!

Next week: Chapter 2

Next meeting: December 14, 2020