http://iseqtools.org/

GATK Best Practices for Variant Discovery

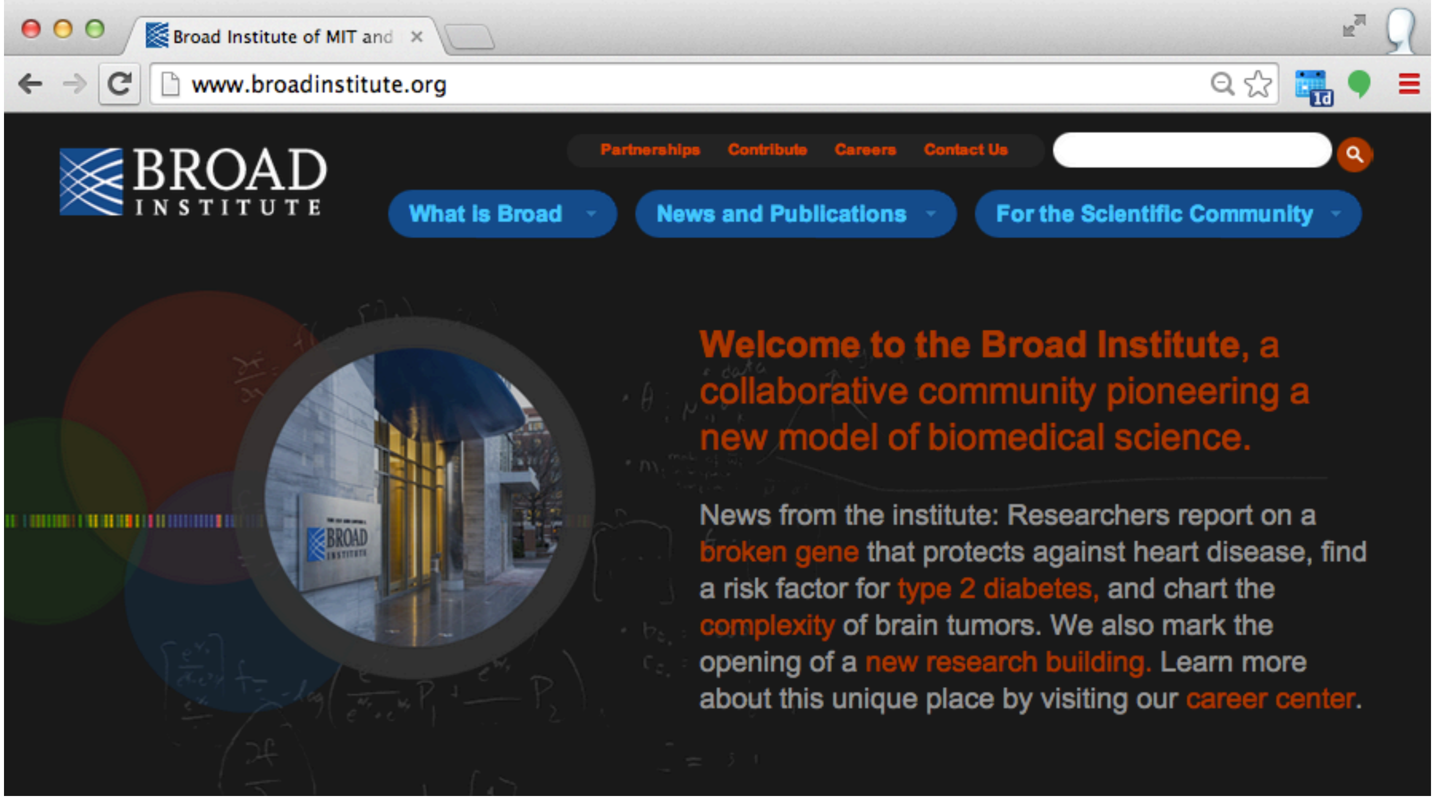
UCLA, Los Angeles CA, USA
2-4 Mar, 2016

Data Sciences & Data Engineering
Broad Institute of Harvard and MIT
http://www.broadinstitute.org/gatk

@gatk_dev

What / who is the Broad Institute?

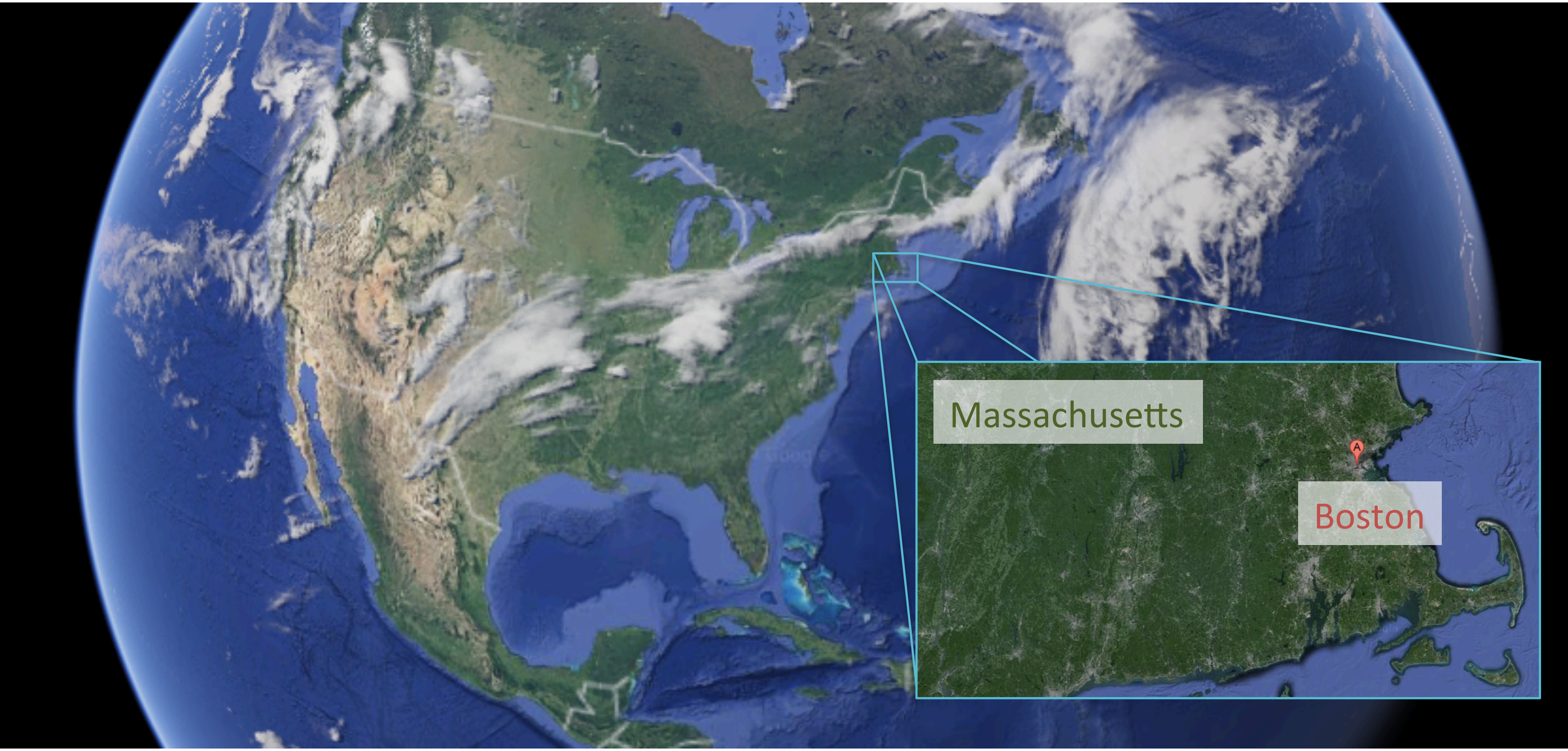
- Spinoff of Harvard & MIT -- Eric Lander and philanthropists Eli & Edyth Broad
- Use the full power of genomics to transform the understanding and treatment of disease



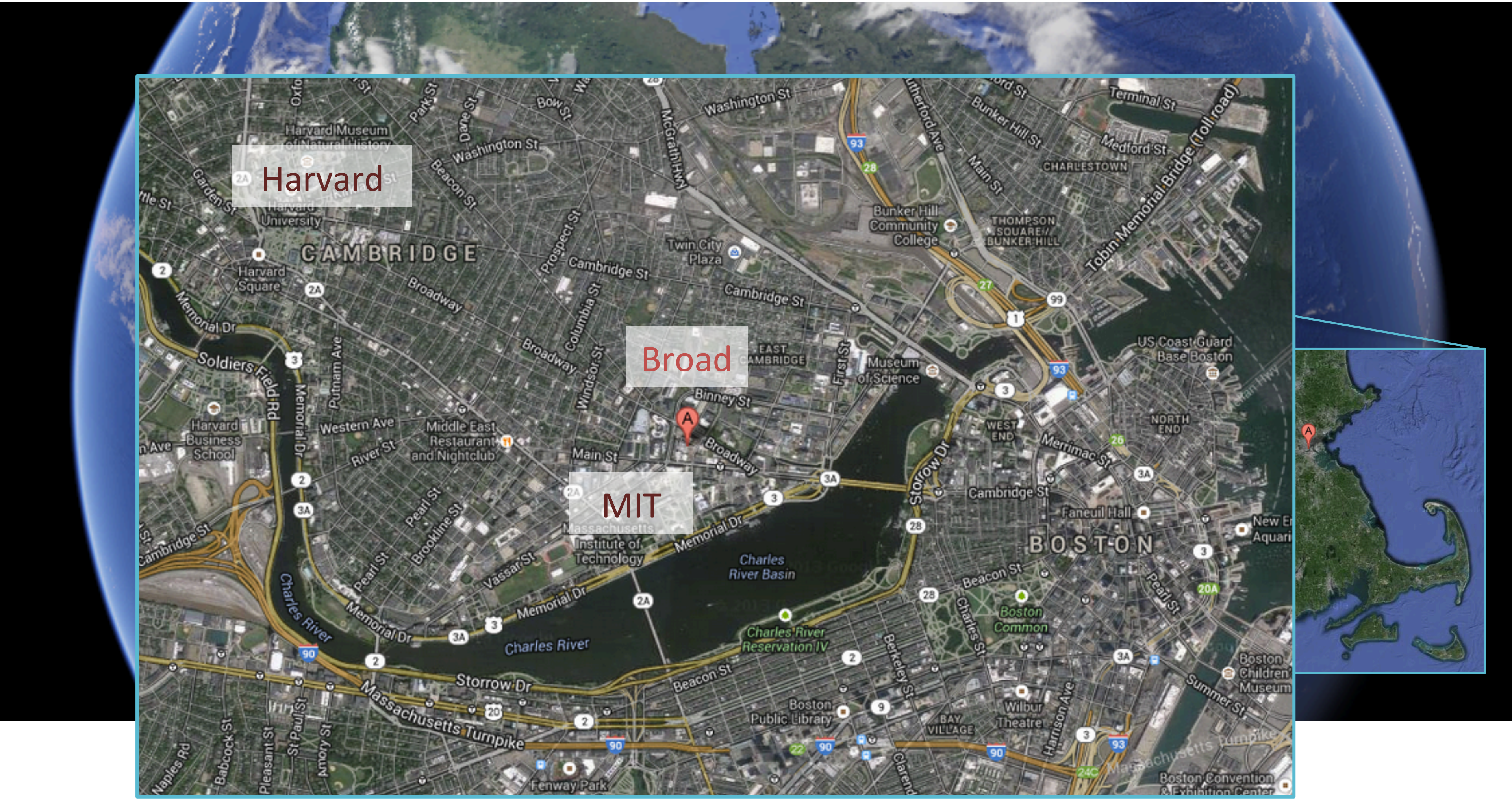
Welcome to the Broad Institute, a collaborative community pioneering a new model of biomedical science.

News from the institute: Researchers report on a broken gene that protects against heart disease, find a risk factor for type 2 diabetes, and chart the complexity of brain tumors. We also mark the opening of a new research building. Learn more about this unique place by visiting our career center.

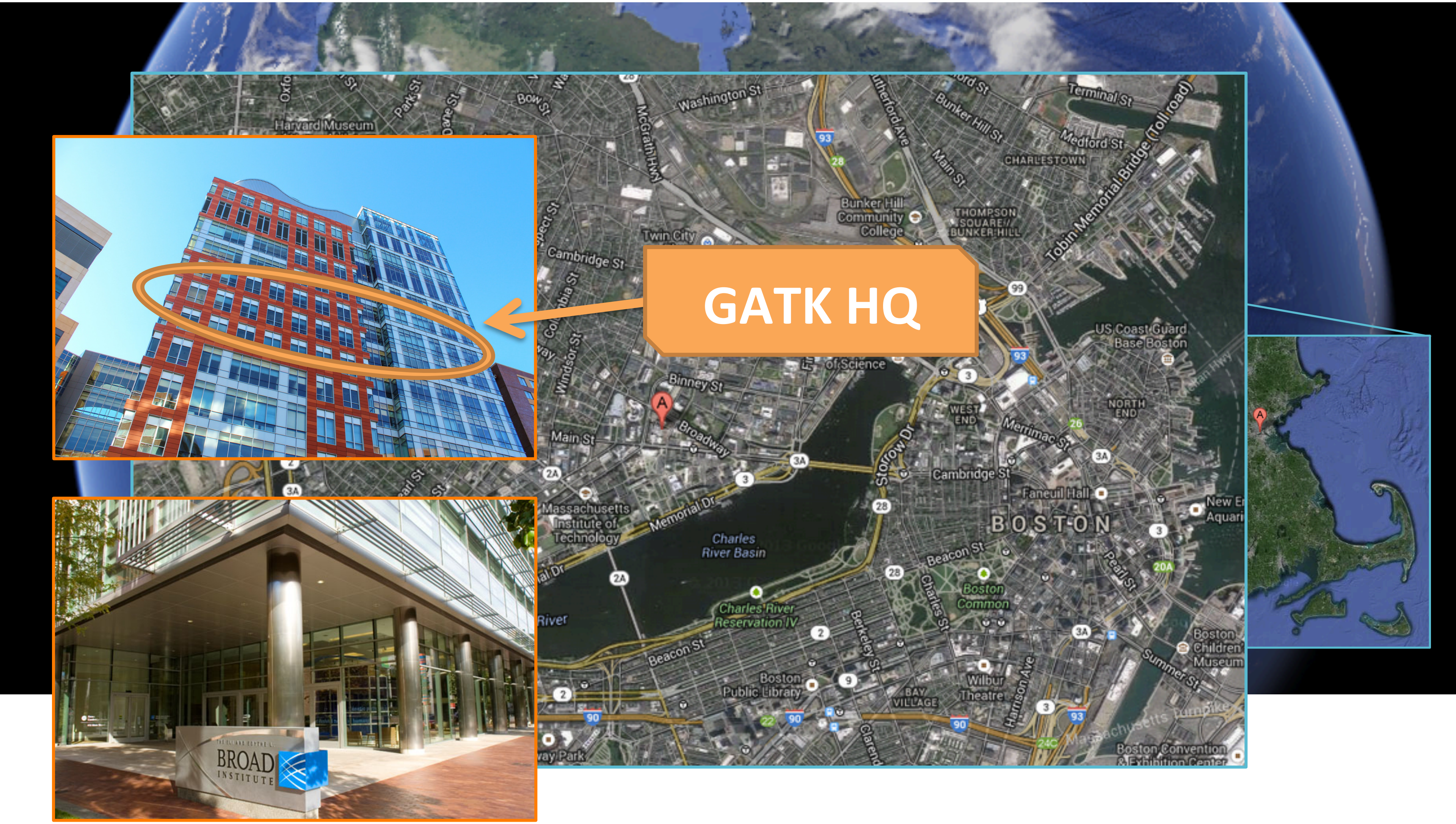
Where in the world is the Broad?



Where in the world is the Broad?



Where in the world is the Broad?



GATK HQ

Data Science & Data Engineering @ Broad

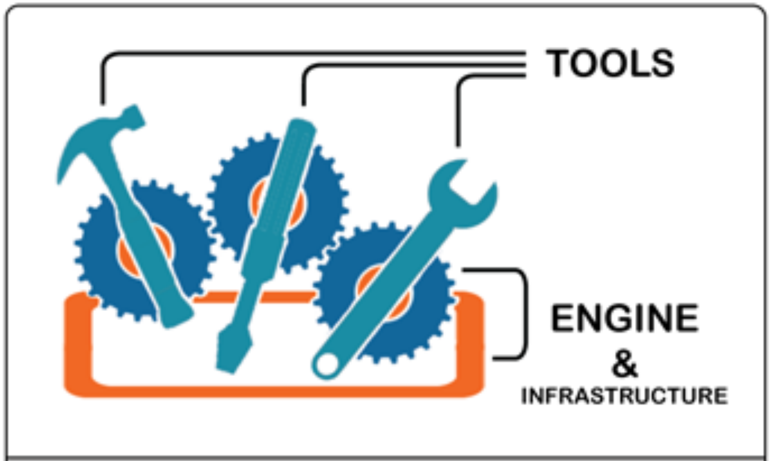
A new organization bringing together software engineers, computational biologists, and computing infrastructure specialists.

A vision that articulates an advanced computing infrastructure, set of data and analysis services leveraging modern cloud computing paradigms.

https://www.broadinstitute.org/dsde/

GATK = Genome Analysis Toolkit

- **Toolkit** focused on variant discovery (SNP & indel)
- **Components:**
  - Engine and infrastructure
  - Tools (walkers)

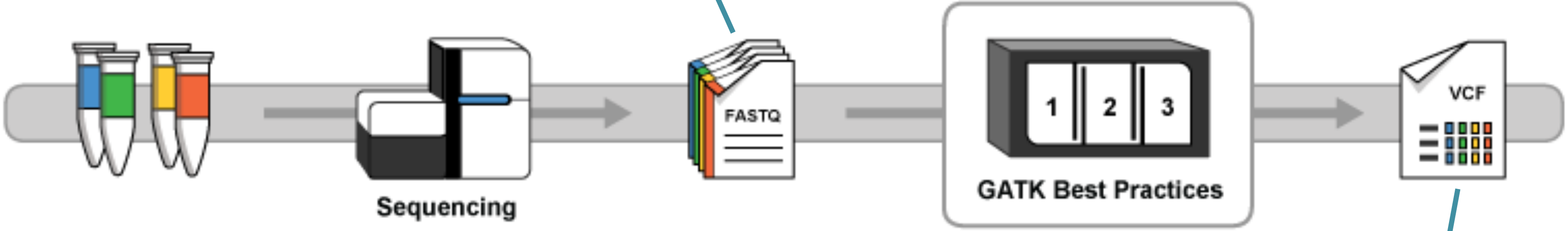


The core GATK components

- Also a **programming framework** for developing genome analysis software

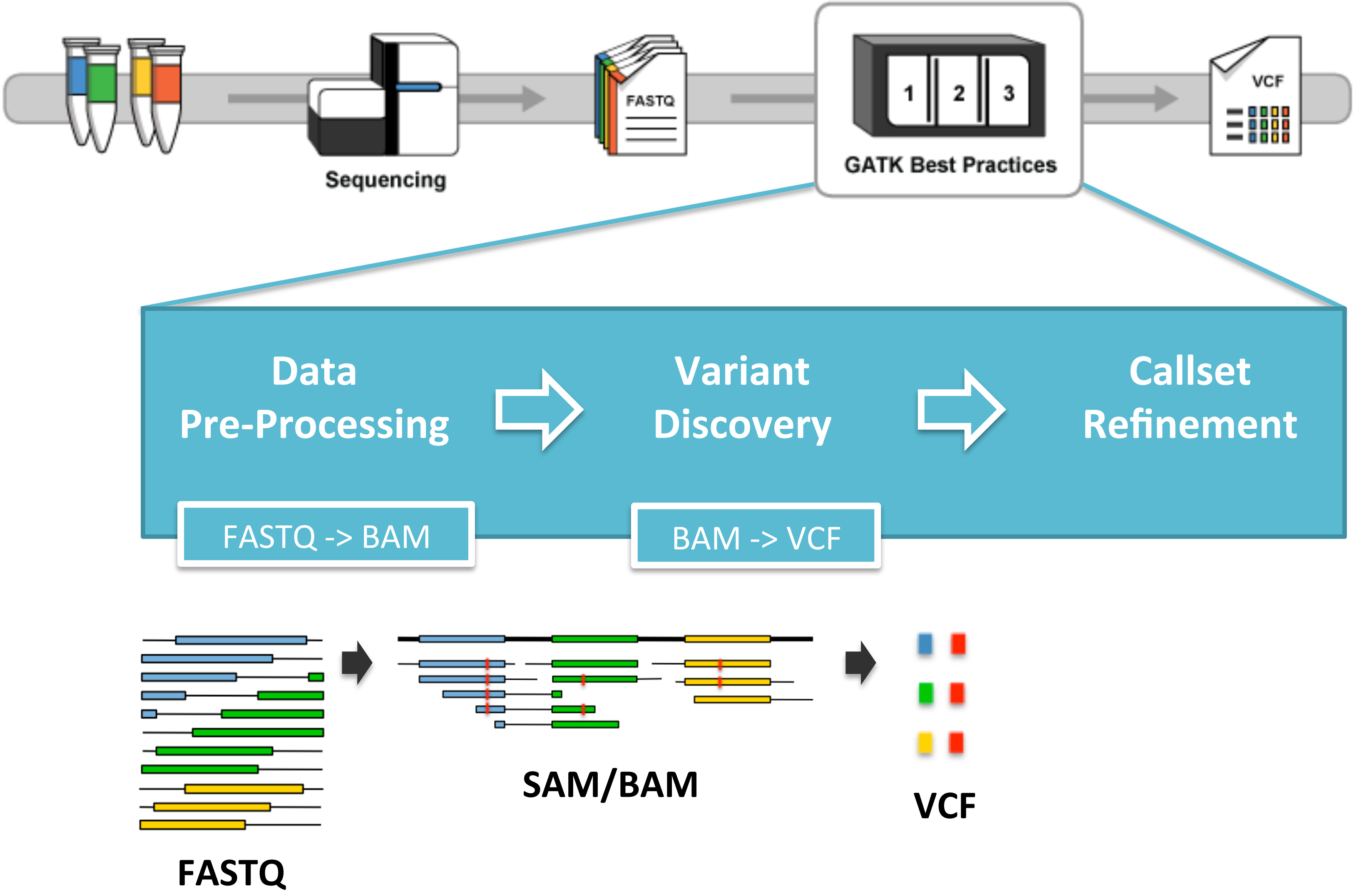
Variant discovery = identify **variants** in sequencing data

RAW READS
SEQUENCE DATA



VARIANTS

GATK Best Practices = complete **reads-to-variants** workflows

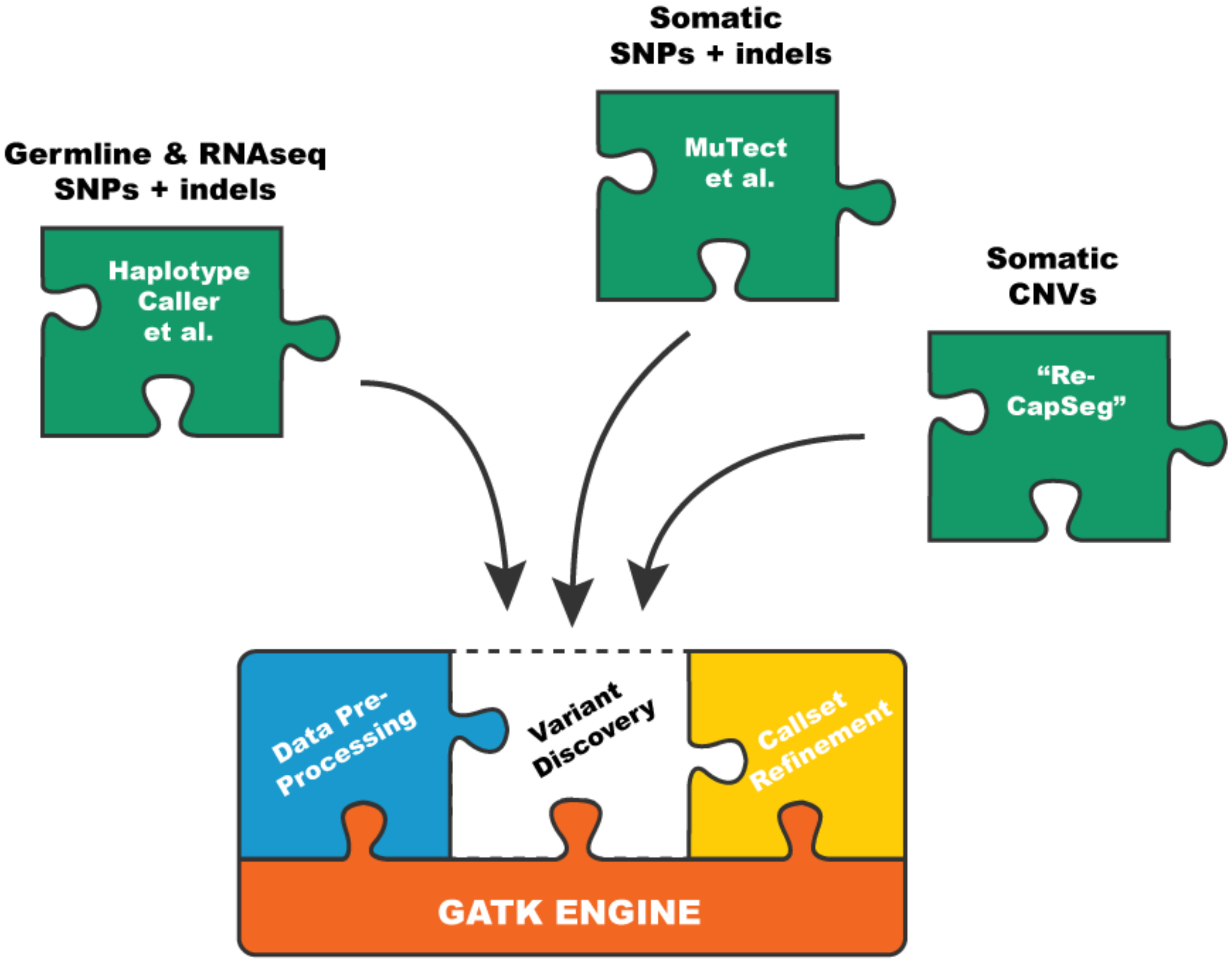


Data Pre-Processing → Variant Discovery → Callset Refinement

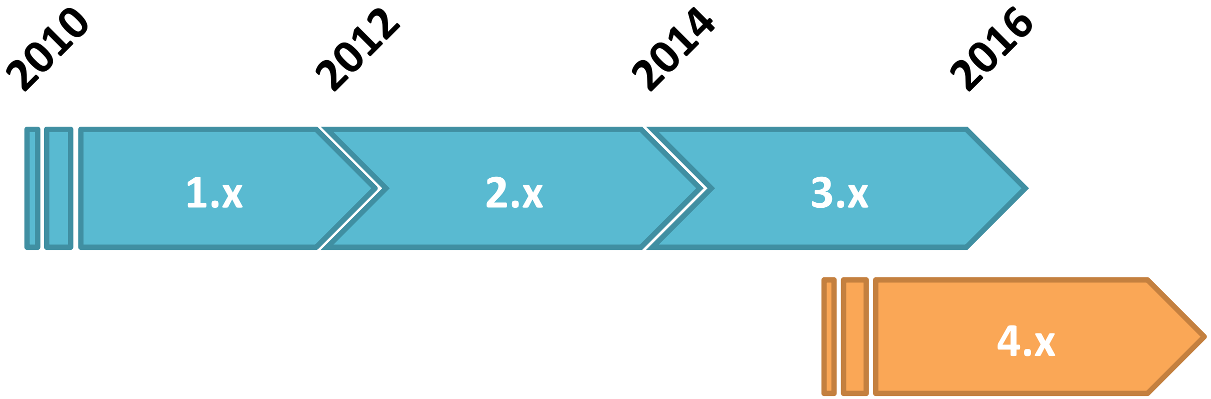
FASTQ -> BAM

BAM -> VCF

Expanding ecosystem of modular Best Practices workflows



GATK development roadmap



Alpha GATK 4 : cloud-friendly and more scalable (Apache Spark)
+ extended functionality (CNVs, Picard)

https://github.com/broadinstitute/gatk

Workshop agenda

Day 1

9 am – 10:30 am
Introduction to Variant Discovery

10:30 am – noon
Pre-processing methods

Day 2

9 am – noon
Germline variant discovery methods

Day 3

9 am – noon
Somatic variant discovery methods

Afternoons

2 pm – 5 pm
Hands-on tutorials