

GATK Workshop



University of Pretoria
South Africa, 22-26 June, 2015

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

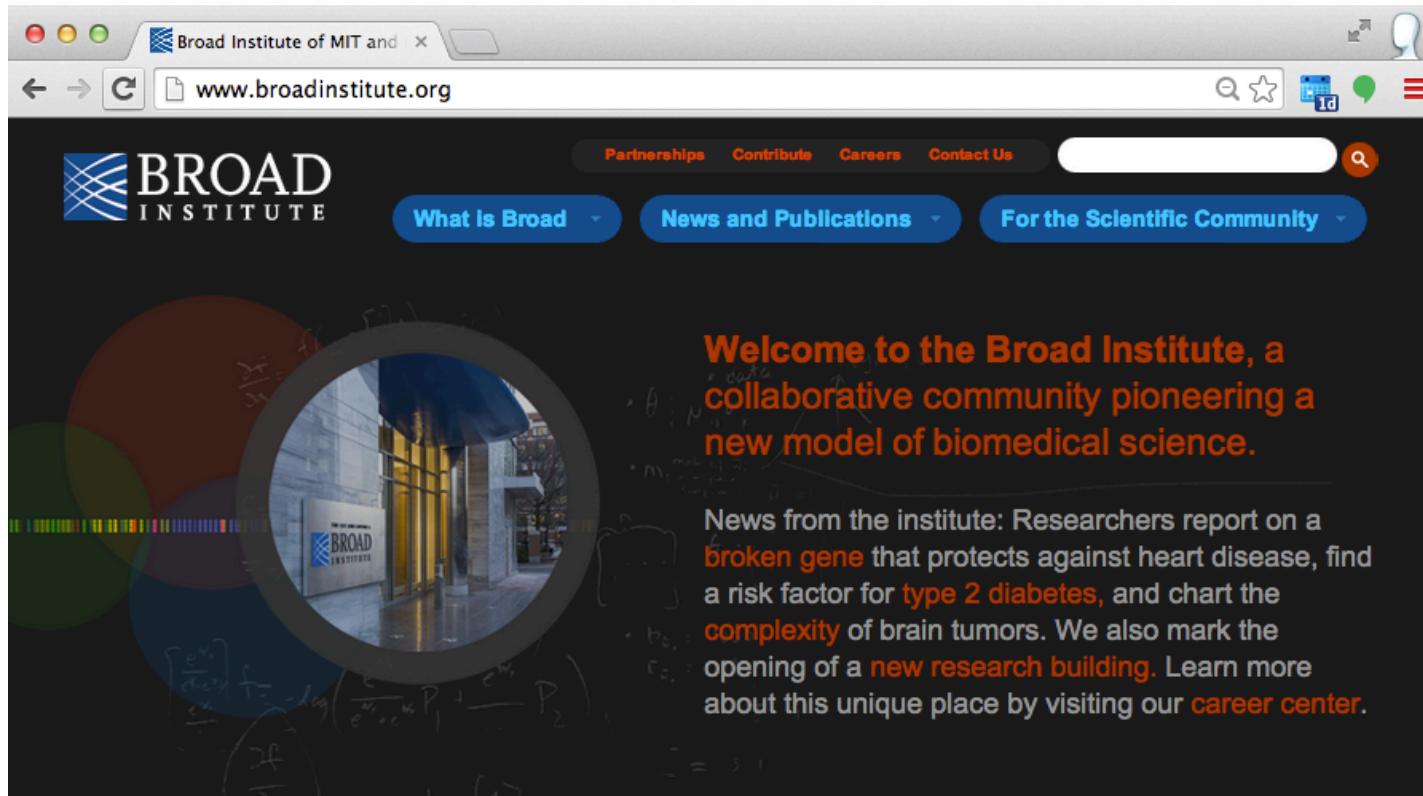


@gatk_dev

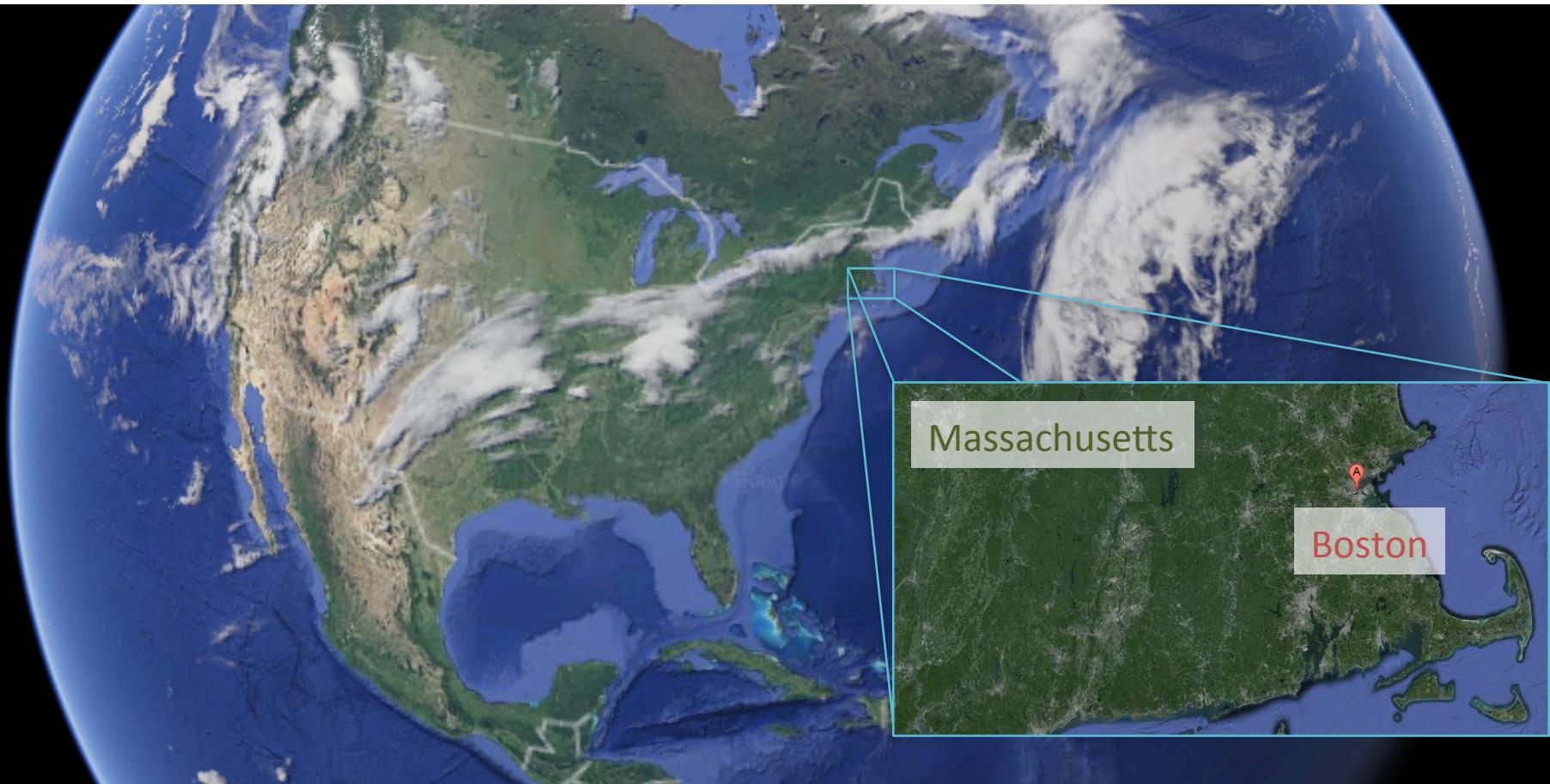


What / who is the Broad Institute?

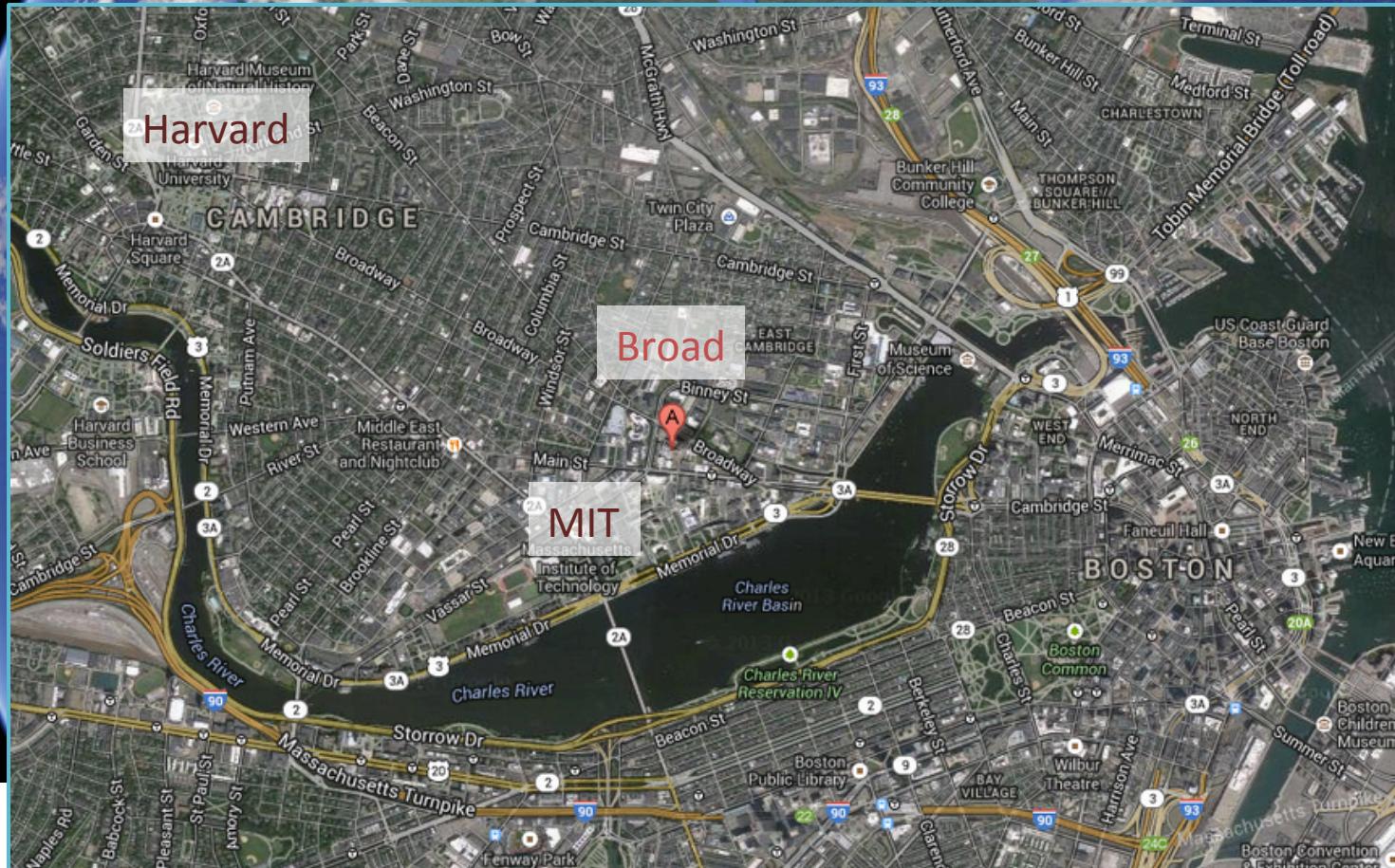
- Spinoff of Harvard & MIT -- Eric Lander and philanthropists Eli & Edyth Broad
- Use the full power of genomics to transform the understanding and treatment of disease



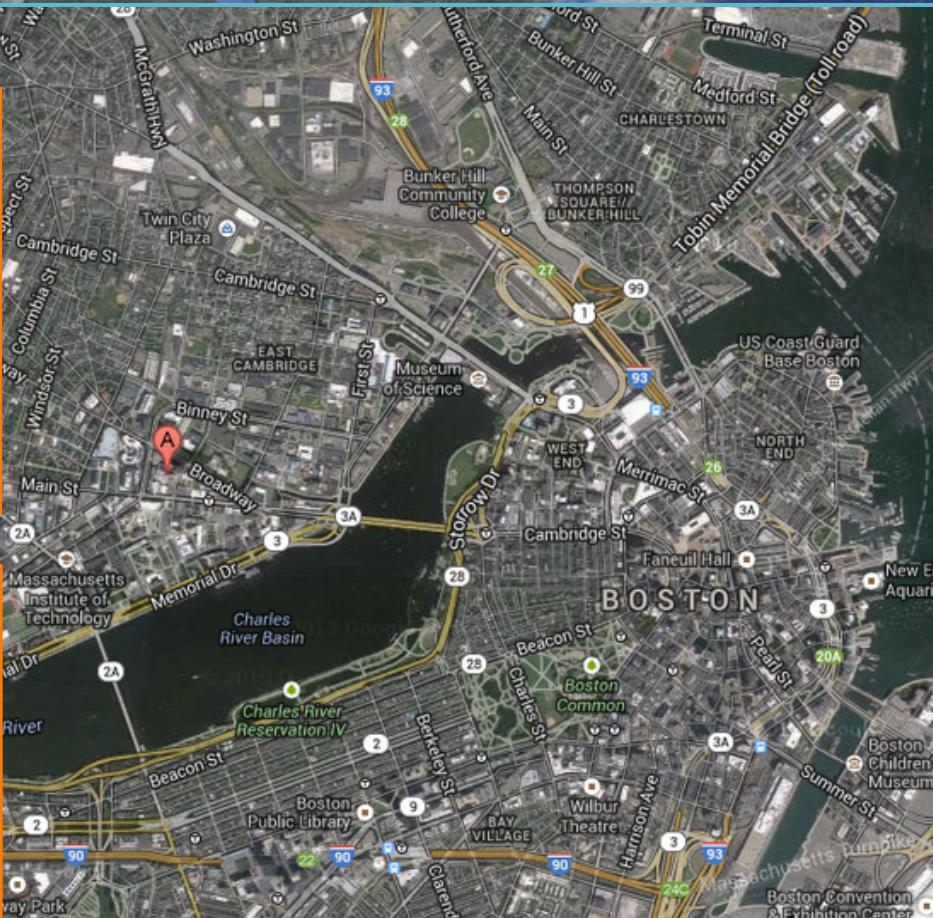
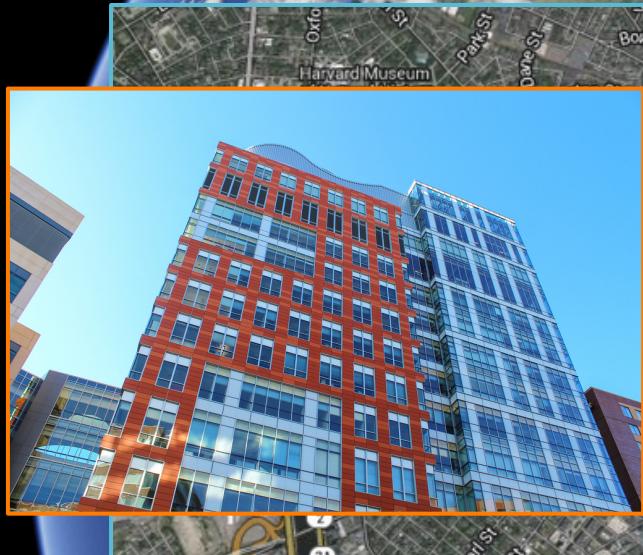
Where in the world is the Broad?



Where in the world is the Broad?



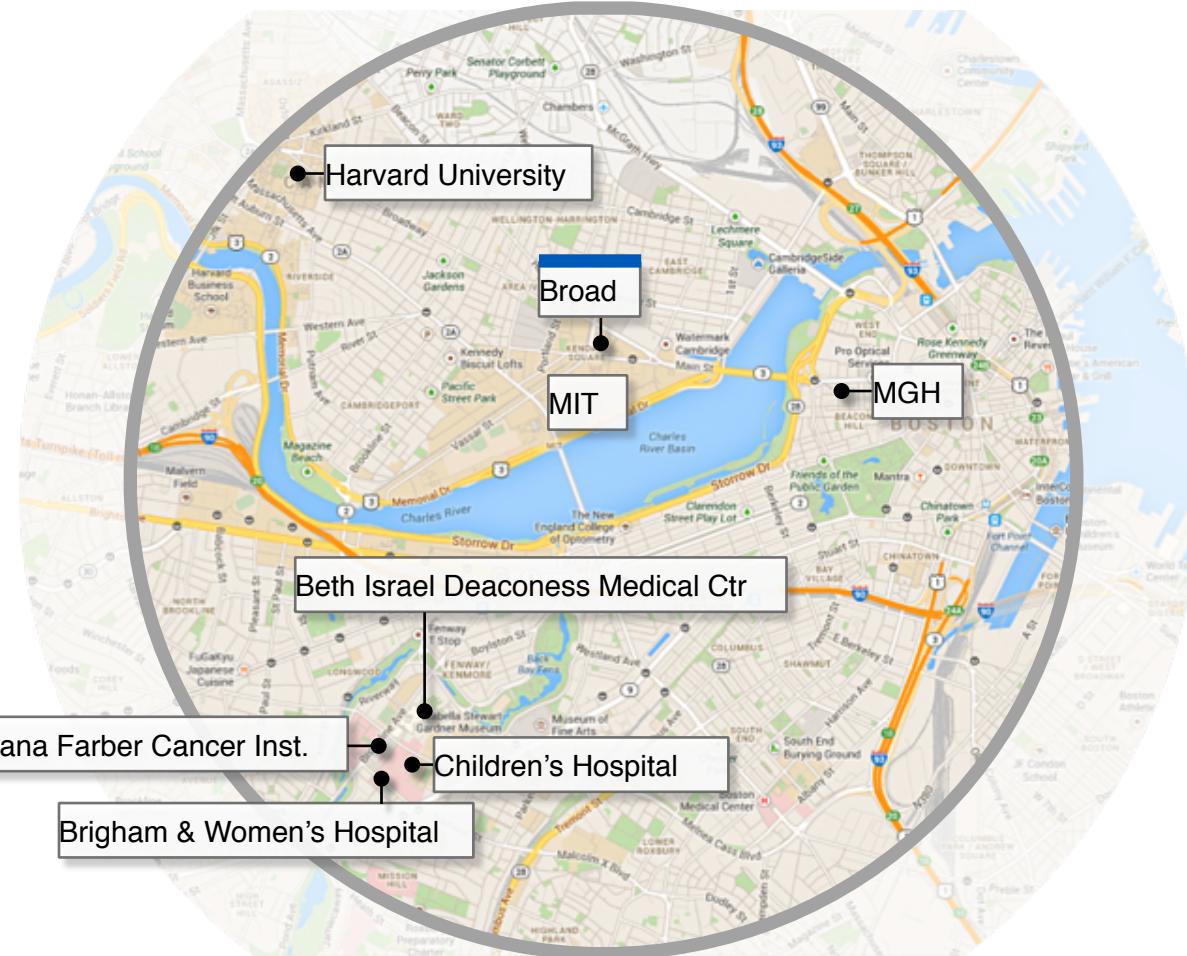
Where in the world is the Broad?



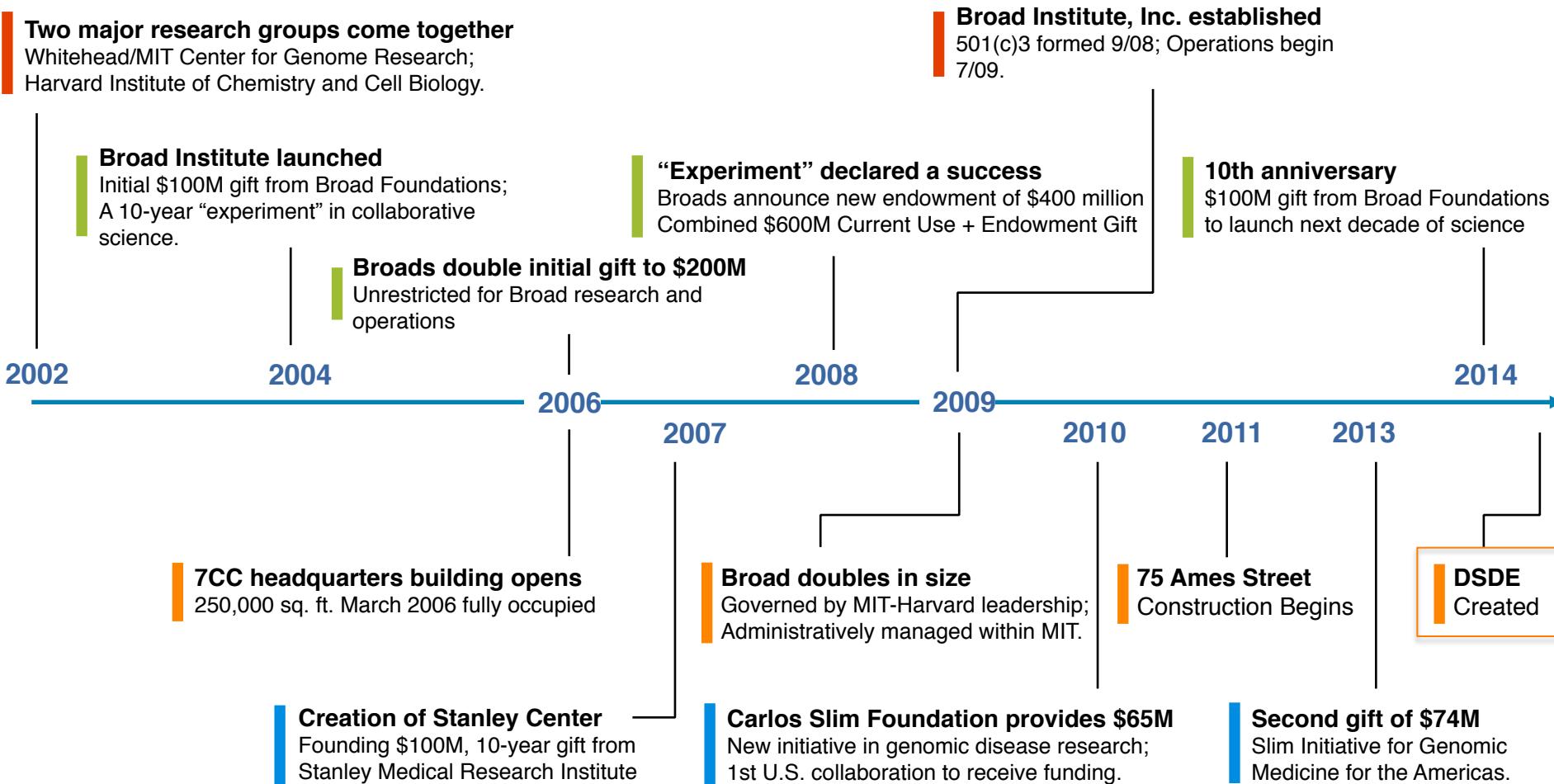
Where in the world is the Broad?



A highly collaborative and research-dense environment



A brief history of the Broad Institute



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/>

DSDE's extended family: GP + CGA + BITS



DSDE: Methods + Engineering + Operations + Communications

Genomics Platform (GP)

Cancer Genome Analysis Group (CGA)

Broad IT (BITS)

+ many collaborators who beta-test the methods and software

General program

Day 1

Introductions, pipelines and QC

AM Lecture session: 9am – 12pm

PM Hands-on tutorial: 1pm – 4pm

Day 2

Data pre-processing (BAM cleanup)

Day 3

Germline variant discovery



Day 4

Somatic variant discovery

Day 5

Working with variants

Day 1: Sequencing for variant analysis

- 9:00 Workshop introduction
- 9:20 High-throughput sequencing for variant analysis
- 10:00 GATK Best Practices and the Broad pipelines
- 10:30 *Coffee break*
- 11:00 Sequence data quality control
- 11:30 Pipelining and parallelism
- 12:00 *Lunch break*
- 13:00 Hands-on session

Approximate schedule – timing may vary depending on Q&A

Day 2: Mapping and cleanup

- 9:00 Recap of Day 1 and Q&A
- 9:15 Mapping and pre-processing
- 10:00 Indel realignment
- 10:30 *Coffee break*
- 11:00 Base recalibration (BQSR)
- 12:00 *Lunch break*
- 13:00 Hands-on session

Approximate schedule – timing may vary depending on Q&A

Day 3: Germline variant discovery

- 9:00 Recap of Day 2 and Q&A
- 9:15 Introduction to joint variant discovery
- 9:45 Variant calling and joint genotyping
- 10:30 *Coffee break*
- 11:00 Variant recalibration (VQSR)
- 11:45 Manual filtration
- 12:00 *Lunch break*
- 13:00 Hands-on session

Approximate schedule – timing may vary depending on Q&A

Day 4: Somatic variant discovery

- 9:00 Recap of Day 3 and Q&A
- 9:15 Introduction to somatic variant discovery
- 10:00 Estimation of contamination with ContEst
- 10:15 *Coffee break*
- 10:45 Variant calling with MuTect
- 11:30 Rescuing TiN variants and eliminating artifacts
- 12:00 *Lunch break*
- 13:00 Hands-on session

Approximate schedule – timing may vary depending on Q&A

Day 5: Somatic variant discovery

- 9:00 Recap of Day 3 and Q&A
- 9:15 Introduction to working with variants
- 9:45 Genotype refinement
- 10:15 *Coffee break*
- 10:45 Evaluating variants
- 11:30 Annotating variants
- 12:00 *Lunch break*
- 13:00 Hands-on session

Approximate schedule – timing may vary depending on Q&A

Documentation is at <https://www.broadinstitute.org/gatk/guide/>

https://www.broadinstitute.org/gatk/guide/tooldocs/

Tool Documentation Index 3.3-0-g37228af

Engine Parameters (available to all tools)

Diagnostics and Quality Control Tools

Sequence Data Processing Tools

Variant Discovery Tools

Variant Evaluation and Manipulation Tools

Help Utilities

Reference Utilities

Validation Utilities

Read Filters

ROD Codecs

Variant Annotations

Current version is

The [Best Practices](#) have been updated for GATK version 3. If you are running an older version, you should seriously consider upgrading. For more details about what has changed in each version, please see the [Version History](#) section. If you cannot upgrade your version of GATK for any reason, please look up the corresponding version of the GuideBook PDF (also in the [Version History](#) section) to ensure that you are using the appropriate recommendations for your version.

https://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNaseq

GATK Best Practices

Recommended workflows for variant analysis with GATK

INDEX DNAseq RNAseq

DNAseq Overview Pre-processing Variant Discovery Suggested Preliminary Analyses

About the DNaseq Variant Analysis workflow

This is our recommended workflow for calling variants in DNaseq data from cohorts of samples, in which steps from data processing up to variant calling are performed per-sample, and subsequent steps are performed jointly on all the individuals in the cohort.

The diagram illustrates the DNaseq Variant Analysis workflow. It starts with 'Raw Reads' which undergo 'Map to Reference (BWA mem)', 'Mark Duplicates & Sort (Picard)', 'Indel Realignment', 'Base Recalibration', and 'Analysis-Ready Reads'. These are then processed by 'Var. Calling HC in ERC mode' to produce 'Genotype Likelihoods'. These likelihoods are used in 'Joint Genotyping' to produce 'Raw Variants (SNPs, Indels)'. These variants are then 'Variant Recalibration separately per variant type' to produce 'Analysis-Ready Variants (SNPs, Indels)'. Finally, these variants are 'Variant Annotation' and 'Phasing' before being evaluated. The evaluation step includes 'Variant Evaluation' with a 'look good?' decision, 'troubleshoot', and 'use in project' options.

See the support forum for questions and bug reports

The screenshot shows a web browser displaying the GATK support forum at gatkforums.broadinstitute.org/categories/ask-the-team. The page title is "Ask the GATK team". A sidebar on the left lists categories like "Recent Discussions", "Activity", "My Bookmarks", "My Discussions", "My Drafts", "Groups", "Participated", "Unanswered", and "Best Of...". Below that is another sidebar with "Categories" and links to "All Categories" (4K), "Social Groups" (0), "Announcements" (120), "Ask the GATK team" (3.2K), "GATK Documentation Guide" (395), "FAQs" (45), "Presentations" (9), "Tutorials" (20), and "Methods and Workflows" (69). The main content area shows a list of questions:

- CalculateGenotypePosterior - supporting file**
Answered ✓ 33 views 7 comments 5 new Most recent by astrand 5:20PM
- haplotyping of side-by-side variants homo/hetero**
Answered ✓ 57 views 4 comments 1 new Most recent by nmbhat 3:22PM
- Picard MergeBamAlignment issue**
Question 17 views 2 comments Most recent by Ryan 1:09PM
- question about ploidy and HC**
Answered ✓ 15 views 3 comments Most recent by Geraldine_VdAuwera 12:32PM
- Only filtering for homozygous SNP's (GATK Unified Genotyper)**
Answered ✓ 32 views 5 comments Most recent by Geraldine_VdAuwera 12:16PM
- Where to obtain HumanNCBI37_UCSC reference sequence?**
Answered 13 views 1 comment Most recent by Geraldine_VdAuwera 11:52AM
- GATK Runtime Error**
Answered ✓ 747 views 11 comments Most recent by ekanterakis 7:57AM
- CombineVariants**
Answered ✓ 17 views 3 comments Most recent by tommycarstensen 5:59AM
- Unified Genotyper settings for 200 half-sibs**
Answered ✓ 15 views 2 comments 1 new Most recent by tommycarstensen 5:18AM
- How can I fix the following error in BQSR**
Answered 20 views 2 comments 1 new Most recent by pkuyh 4:06AM

General program

Day 1

Introductions, pipelines and QC

AM Lecture session: 9am – 12pm

PM Hands-on tutorial: 1pm – 4pm

Day 2

Data pre-processing (BAM cleanup)

Day 3

Germline variant discovery



Day 4

Somatic variant discovery

Day 5

Working with variants