

BioCompute: Standard to Communicate Bioinformatic Workflow Information and Ease Organizational Burden

Jonathon Keeney, Ph.D.
Assistant Research Professor
The George Washington University

Hadley King
hadley_king@gwu.edu

Janisha Patel
janishapatel@gwmail.gwu.edu

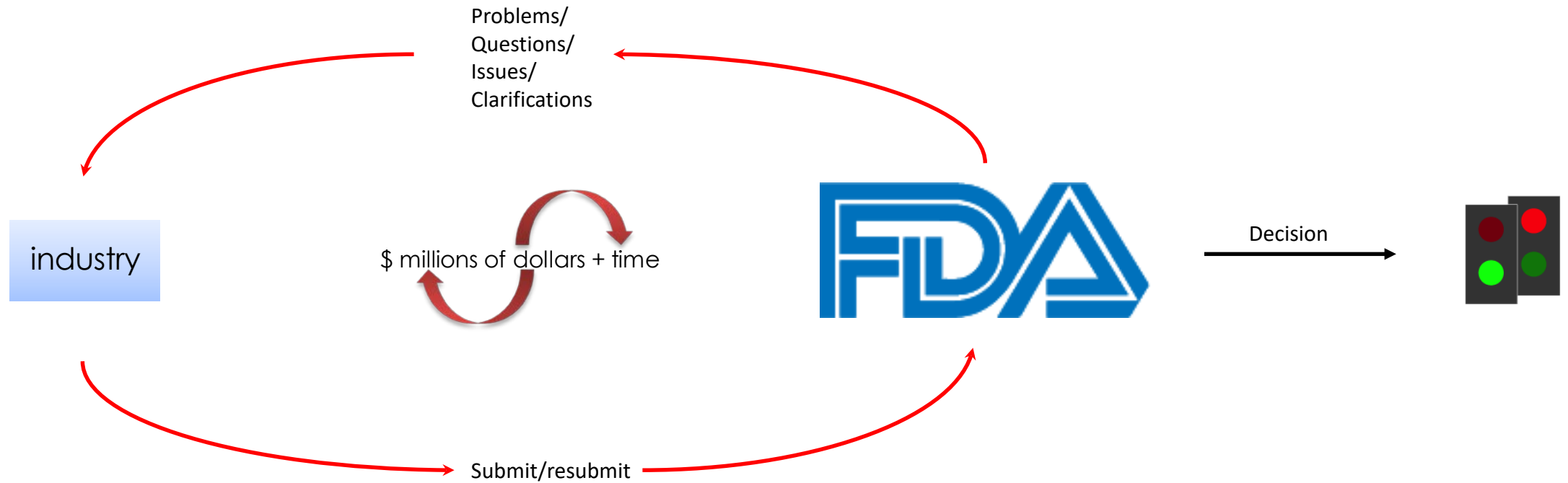
Guidelines

1. Please turn off video
2. Please mute
3. Unmute for questions or post in chatbox
4. Handouts that will be referenced will also be provided via chat
5. Please use Chrome to create your own BCO during the demo exercise

Goals of This Workshop

1. Present BioCompute to anyone new to the project or hasn't kept up with it lately
2. Demonstrate our format for future workshops for FDA personnel
3. Get feedback
 - Possible workflows that FDA personnel may find relevant
 - Best practices

Wasted Time and Money



A solution should...

- Be **human readable**: like a GenBank sequence record
- Be **machine readable**: structured information with predefined fields and associated meanings of values
- Contain enough information to understand the computational pipelines, interpret information, maintain records, and reproduce experiments
- Have a way to be sure the information has not been altered: immutable

Solution: BioCompute



- Acts like an envelope for entire pipeline
 - Can incorporate other standards (e.g. CWL)
- Built in collaboration with the FDA
- Human and machine readable
 - Written in JSON
- Categorized by domains
- Adheres to and encourages F.A.I.R. principles
 - Fully open source
- Adaptable
 - e.g. to other schemas
- Preserves data provenance
- Unique IDs for versioning
- IEEE approved Standard for communicating genomic analysis workflows

802.11 Analogy





BioCompute Object

Top Level

BCO ID: <https://w3id.org/biocompute/1.3.0/examples/FDA-NA-TestsBreastCancer>
Checksum: 06DACE70679F35BA87A3DD6FFFED4ED24A4F5B8C2571264C37E5F1B3ADE04A31
Specification: <https://w3id.org/biocompute/1.3.0/>

Metadata

Provenance Domain

Name: FDA-NA-TestsBreastCancer
Version: 1.0
Review:
approved: Natalie Abrams, NIH ; createdBy
Created: 2018-05-24T09:40:17-0500
Modified: 2018-06-21T14:06:14-0400
Embargo: Start: 2000-09-26T14:43:43-0400 End: 2000-09-26T14:43:45-0400
Contributors:
Janisha Patel (<http://orcid.org/0000-0002-8824-4637>), George Washington University; createdBy, modifiedBy
Dara Baker, George Washington University; authoredBy
License: <https://spdx.org/licenses/CC-BY-4.0.html> --> licensing is inferred by OncoMX licensing. Pub=

Parametric
domain

Usability Domain

FDA-approved or cleared nucleic acid-based human biomarker tests for breast cancer
The .xlsx file FDA-NA-TestsBreastCancer.xlsx contains FDA-approved human biomarker tests for breast cancer.
Each row represents one gene linked to its respective test. Genes are identified by UniProtKB, HgncName, EDNR number
Tests are distinguished by manufacturer, FDA submission ID(s), clinical trial ID(s) and PubMed ID(s).

Usability domain

Extension Domain

Dataset Extension:
Comment: Unique column headers for the dataset
Test_disease_use: FDA-listed disease corresponding to approved test
test_trade_name: FDA-listed product name
test_manufacturerfee: FDA-listed patent company for the approved test
sest_submission: FDA submission ID(s), web links; FDA-listed patent ID associated with test
test_is_panel: A single biomarker or biomarker panel? Y for yes, N for no
gene_symbol: HGNC ID from <https://www.genenames.org>
uniprotKB_ac: UniProtKB from <https://www.uniprot.org>
biomarker_id: Matched to EDNR IDs based on HGNC Name
biomarker_origin: Characteristic that makes this a biomarker; molecular abnormalities that can lead to cancer
ncit_biomarker: Searchable terms for gene/Biomarker from NCI Thesaurus (NCIt)

Extension
domain

Description Domain

Keywords: cancer, breast cancer, biomarker, biomarker test, FDA, UniProtKB, EDNR
External References: (Name, Namespace, Ids)
PubMed; pubmed;
UniProt; accession;
EDNR; EDNR number;
HGNC; HgncName;
GTR; GTR terms;
Platform: Manual
Pipeline Steps:
Step 1: Download FDA-approved tests
Description: FDA-approved tests were downloaded a list of FDA-approved or cleared nucleic acid based tests
Input List: <https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm>
Output List: ~/FDA-approved-or-cleared-NA-based-tests

Description
domain

Execution Domain

Scripts: none
Script Driver: manual
Software Prerequisites: None
External Data Endpoints:
Name In Vitro Diagnostics > Nucleic Acid Based Tests
URL <https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm>
Name NCBI Genetic Testing Registry
URL <https://www.ncbi.nlm.nih.gov/gtr/>
Environment Variables: None

Execution domain

Parametric Domain

N/A

Parametric domain

Input/Output Domain

Input Subdomain:
Filename: Multiple test files from "Nucleic Acid Based Tests: List of Human Tests"
Access Time: 2018-10-10T11:34:02-5:00
URI: <https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm>
Output Subdomain:
Filename: FDA-NA-TestsBreastCancer.xlsx
Media Type: xlsx/csv
Access Time: 2018-10-10T11:37:02-5:00
URI: <https://docs.google.com/spreadsheets/d/1xUY7WJNEZHyCgH5sYpxEuqAbtgVUuWgR2oc0IwhH28Y/edit#gid=1492026303>

IO
domain

Error Domain

Error domain

ACCESS: Private | NAME: test-workflow | ORG: dnanexus.science | ADDED BY: sam.westreich | ID: workflow-FQ7P7Vj05922F6k6J3b87yQ6

CREATED: 2018-12-10 23:16:23

Edit tags

Revision: 1 | Latest | Edit | Fork | Export | Run Workflow rev1

SPEC | WORKFLOW DIAGRAM

INPUTS

file	Input 1	REQUIRED	workflow-app-1
file	Input 2	REQUIRED	workflow-app-2

OUTPUTS

file	Output 1	REQUIRED	workflow-app-1
file	Output 2	REQUIRED	workflow-app-2



Projects | Data | Apps

Identifiers and File name(s) | Search | Queries | Save Query | Copy files to project

Start Query From:

- Case
- File
- Sample
- Portion
- Slide
- Analyte
- Aliquot
- Drug therapy
- Radiation therapy
- Follow up
- New Tumor Event

Workflow diagram showing steps: File (ADD FILTER) -> Data Format (Remove filters) -> Experimental Strategy (Remove filters) -> Disease Type (Remove filters)



Galaxy Administration

Galaxy Administration

Galaxy Administration | Analyze Data | Workflow | Shared Data | Admin | Help | User | Using 35.2

Administration

- Security
 - Manage users
 - Manage groups
 - Manage roles
- Data
 - Manage quotas
 - Manage data libraries
- Server
 - Reload a tool's configuration
 - Profile memory usage
 - Manage jobs
 - Manage installed tool shed repositories
- Tool sheds
 - Search and browse tool sheds
- Form Definitions
 - Manage form definitions
- Sample Tracking
 - Manage sequencers and external services
 - Manage request types
 - Sequencing requests
 - Find samples

Repository Actions | Tool Shed Actions

Genome/Exome paired analysis (SNVMix1)

Boxes are red when tools are not available in this repository (this page displays SVG graphics)

Main | Home | HIVE Portal | Links

CensusScope

HMB25-2_R1

Parameters

Progress

Results

Taxonomy Details

Taxonomy Help

Convergence

Phylogenetic Tree

Tree Tree

Table

Subhost

What's Next?

Alignment

Loading Status

Task	Progress
Building histogram	Done 100%
Preparing alignments	Done 100%
Visualizing alignments in track	Done 100%
Fetching alignments	Done 100%
Creating rotation tree diagram	Done 100%

Taxonomy Help | Taxonomy Details

Alcemy

SubprojectID

Name

Taxname

Parent

Rank

Taxonomy ID



BCO Portal

BioCompute Editor



Sign in

SIGN IN NOW

Don't have an account? [Sign up](#)

<https://portal.aws.biochemistry.gwu.edu/sign-in>

BioCompute Object (BCO) App-a-thon

May 14 through October 18



Integrating with Other Standards

- Institute of Electrical and Electronics Engineers Standard (approved January 2020)
- International Standards Organization certification expected by Q4 2020 through joint agreement



Major Changes to IEEE Version

- **Hosted on open source repo**
 - Hosted on GitLab – can support independent branches and public comments
- **Explicit call to JSON Schema version**
 - Allows interaction with other standards (e.g. W3CProv)
 - Explicitly indicate “required” fields
- **Md5 no longer used for hashing**
 - “e-TAG” ensures object is not modified after submission
- **Full version can be seen here:**
 - <https://gitlab.com/IEEE-SA/2791/ieee-2791-schema>
 - This is Version 1.4 according to internal numbering. Previous (unstandardized) version is 1.3

BCO Timeline

1st BioCompute
Workshop
March, 2014

2014

2015

2016

2017

2018

2019

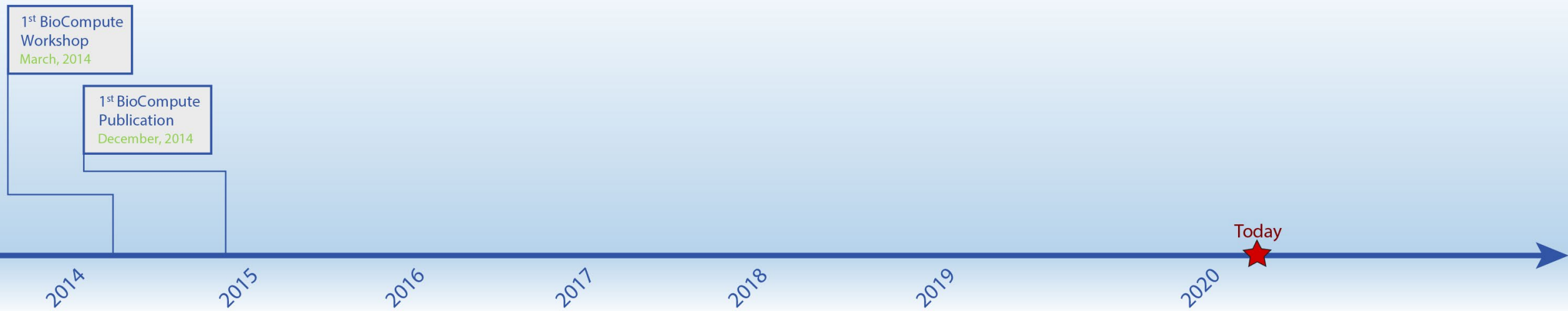
2020

Today

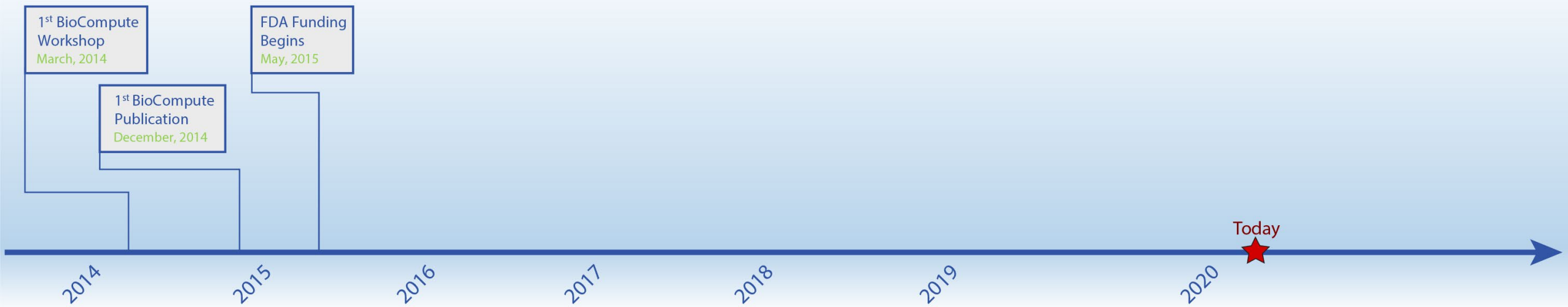


BioCompute
Objects

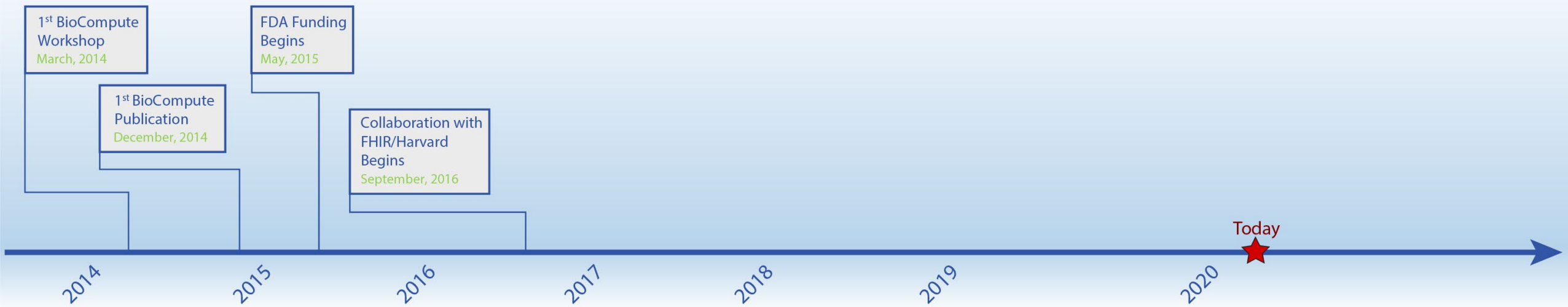
BCO Timeline



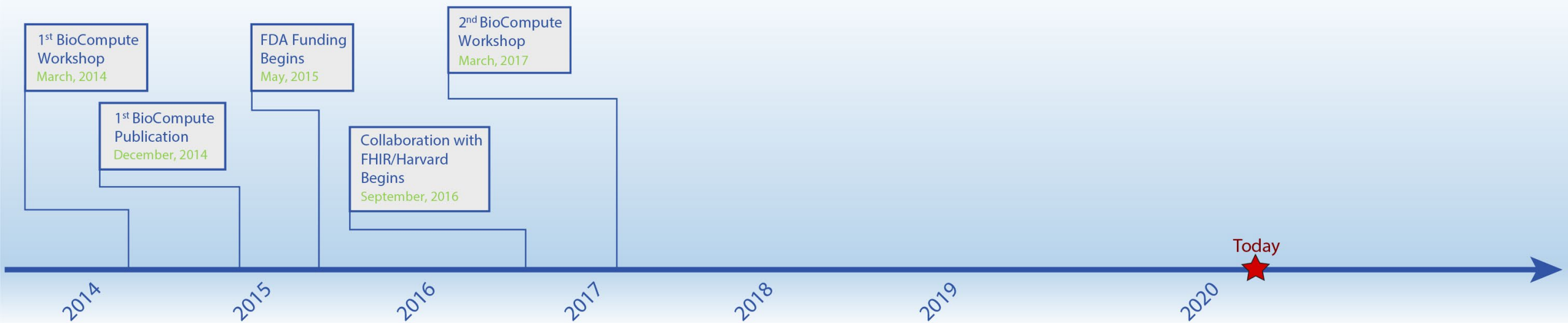
BCO Timeline



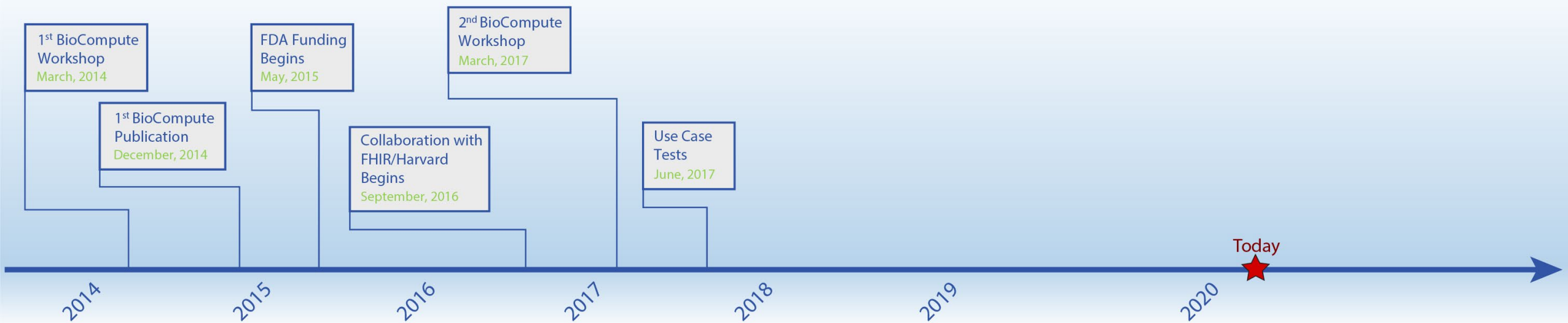
BCO Timeline



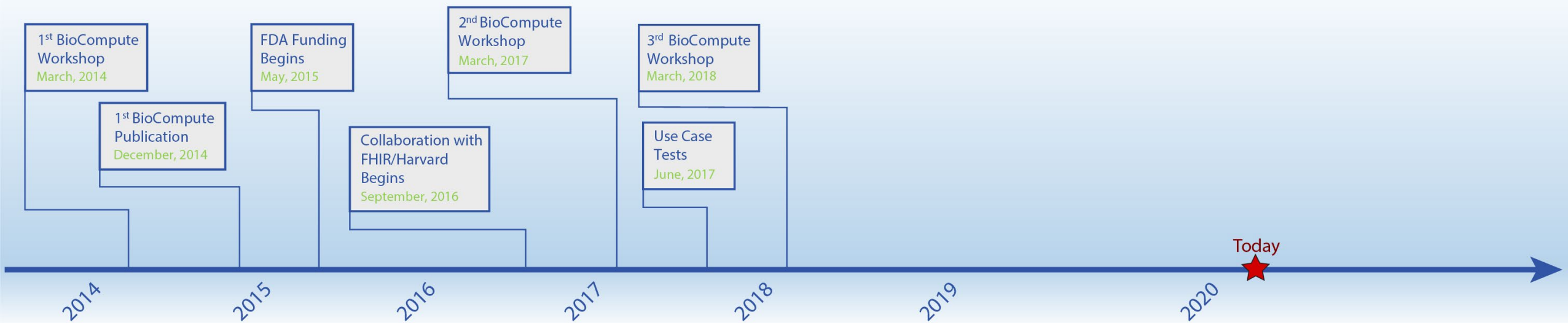
BCO Timeline



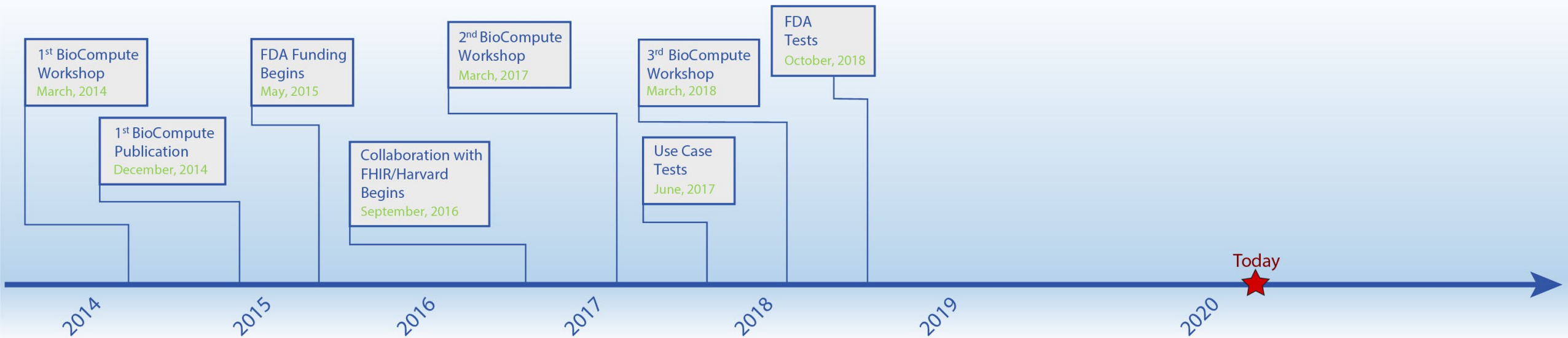
BCO Timeline



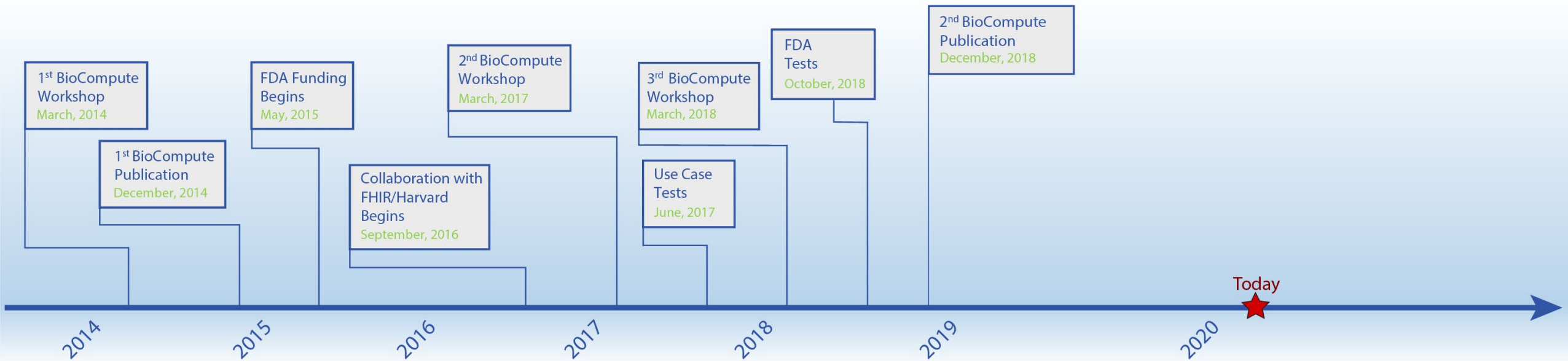
BCO Timeline



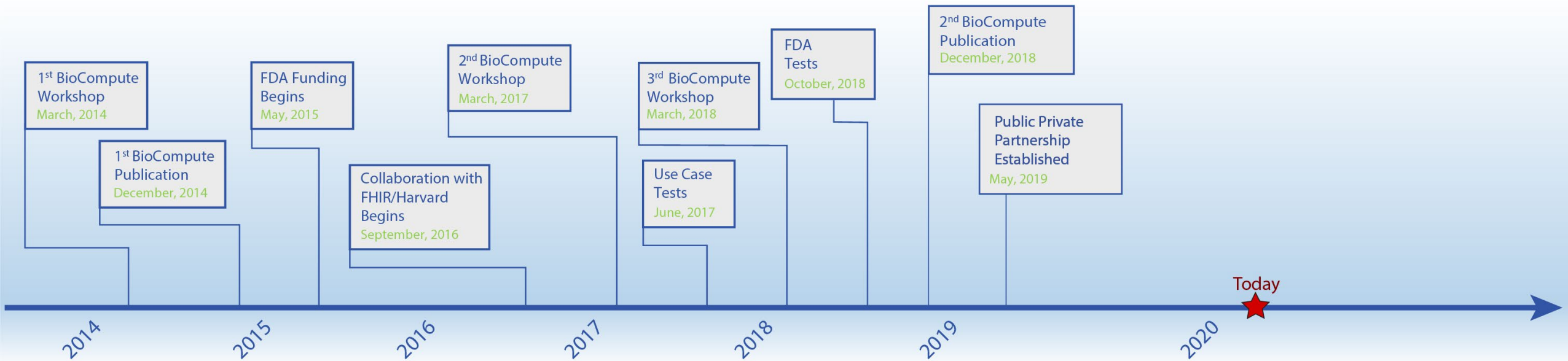
BCO Timeline



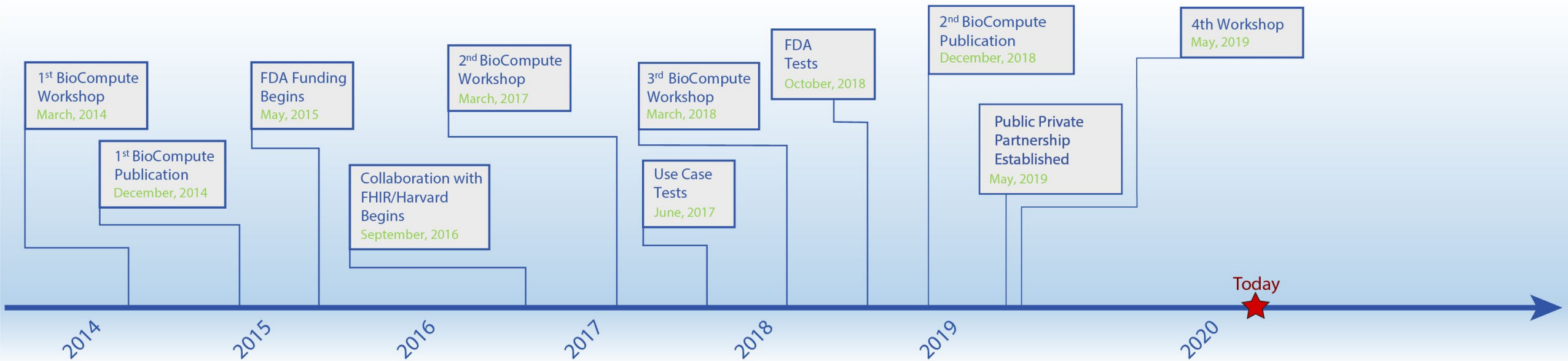
BCO Timeline



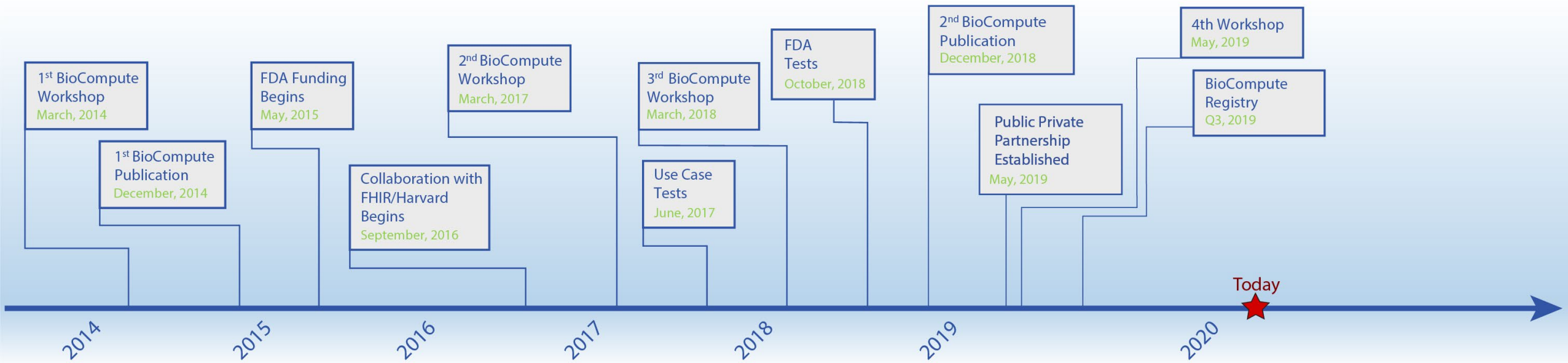
BCO Timeline



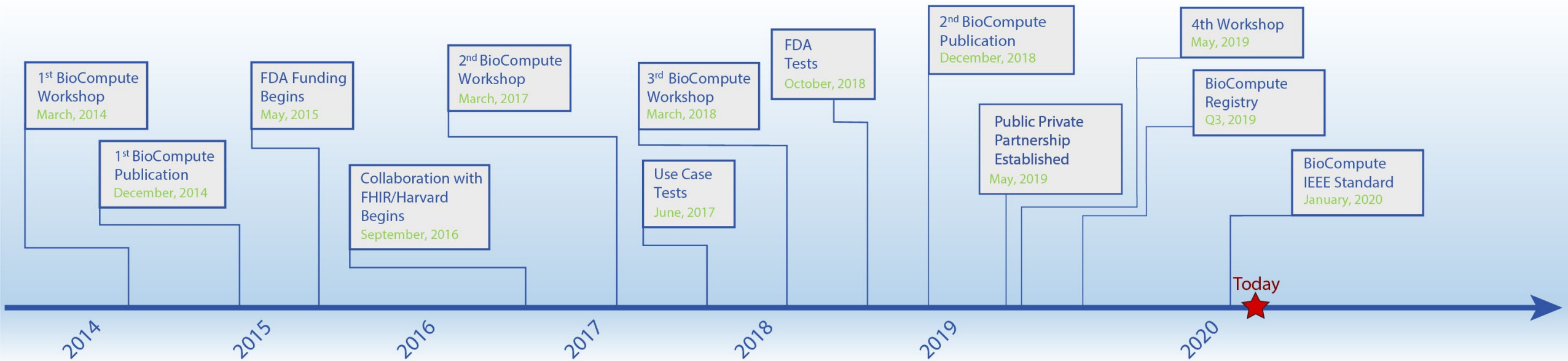
BCO Timeline



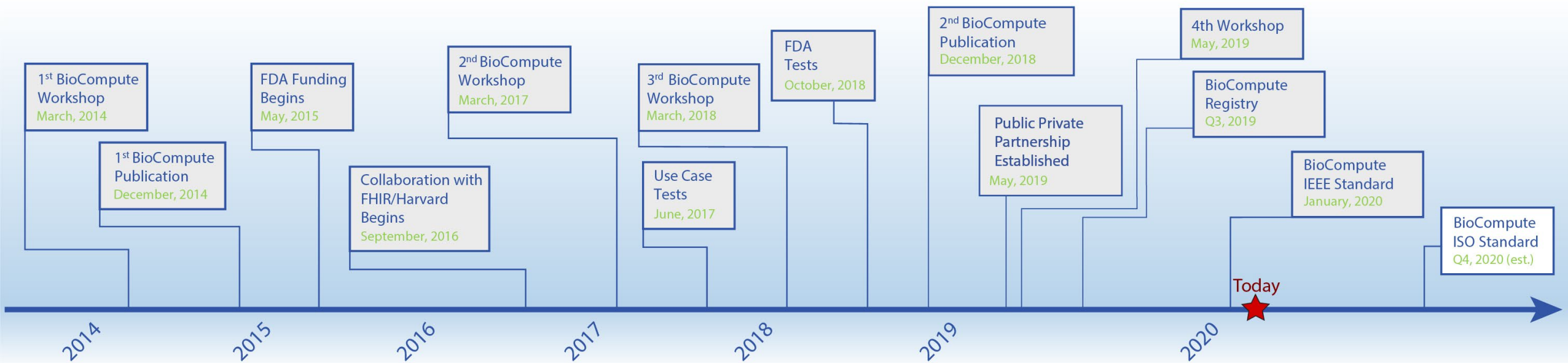
BCO Timeline



BCO Timeline



BCO Timeline



BioCompute Advisory Boards Workshop

Introductions	2:00 – 2:10PM	Use Case Selection	2:55 – 3:00PM
All		All	
Introduction to BioCompute	2:10 – 2:20PM	BCO Editor Tutorial	3:00 – 3:05PM
Jonathon Keeney		Janisha Patel	
Advisory Boards' Aims	2:20 – 2:30PM	BCO Creation	3:05 – 3:55PM
Hadley King, Jonathon Keeney		All	
Use Case Brainstorming	2:40 – 3:10PM	Closing Remarks	3:55 – 4:00PM
Hadley King: Command line example (5 min) Janisha Patel, Platform example (5 min) Discussion (20 min)		Jonathon Keeney	

Regulatory Advisory Board (RAB)

- **AIM 2: Develop a mechanism for transfer of BCOs**
 - **Subaim 2.1 Determine, document and implement security for BCO transfer.**

To ensure proper security implementation of BCO transfer, an FDA Regulatory Advisory Board (RAB) of policy experts will be created to determine acceptable reference information criteria. During year one, the RAB and our team will work to develop an action plan on how the guidance and recommendations of RAB will be implemented throughout the project. At the close of year two, a “Best Practices” document detailing the RAB’s guidance will be posted via GitHub (or another similar public forum). In year three, the RAB will assist in the development of recommendations on how to best utilize Drug Master File (DMF) submissions within the BCO framework.

Technical Advisory Board (TAB)

- **AIM 1: Develop a BioCompute database (BioComputeDB)**
 - **Subaim 1.1 Host informational training meetings and use-case collection meetings across centers to obtain center specific BCO needs.** In year 1, we will develop FDA Technical Advisory Board comprised of technical experts to determine content areas of BCOs for initial focus. The advisory board will also provide suggestions for sponsors who wish to share information or participate in the BioComputeDB development.

Discussion: Feedback

- The way that information is captured will depend on the environment the analysis is run in. As a Reviewer, what is the best format for representing file structure?

What are the “best practices?”

- E.g. for a spike-in study with multiple versions of the same pipeline, do you prefer multiple BCOs that reference each other? Or a single BCO?

HIVE Platform Example
Manual QC step: Usability?

Command Line Example
How are files represented?

<https://hive.biochemistry.gwu.edu/confluence/display/BUW/BioCompute+Workshop>

Thank you!

- Your time and feedback are greatly appreciated!
- Project specific feedback will be hosted here:
 - <https://hive.biochemistry.gwu.edu/confluence/display/BUW/BioCompute+Workshop>