Firehose in The Cancer Genome Atlas: Rigorous Open Science, At Scale

Michael S. Noble and The Broad Institute GDAC Team
Broad Institute of MIT and Harvard
mnoble@broadinstitute.org
Site URL: http://gdac.broadinstitute.org
Source URL: not applicable
Open Source License: not applicable

The Cancer Genome Atlas (TCGA) has generated an unprecedented wealth of biomedical information: with over a petabyte of primary sequence, clinical, and characterization data generated from more than 10,000 patients, in 11 distinct modalities for 30 cancer types. This magnitude and multidimensional heterogeneity presents a number of challenges to cancer researchers: traditional approaches to data management, curation and computational pipelining do not scale; analysis algorithms undergo constant evolution as scientists grapple with making sense of the underlying biology in the data; and to derive high-level insight these data and algorithms must be integrated within complex workflows that go beyond the expertise or staffing levels of any single individual, group, or institution.

The Firehose analysis pipeline at the Broad Institute TCGA Genome Data Analysis Center (GDAC) was born to address these challenges: by regularly aggregating and normalizing data into versioned packages; automatically performing a suite of cutting-edge, integrative bioinformatic analyses upon them; and making the results of such publicly available, en masse through `firehose_get` and online in the form of biologist-friendly reports and APIs. Thousands of analyses and reports are generated every month, each similar to the Results section of a publication but without the delay of peer-review, and tagged with a persistent DOI for citation in the literature.

By applying large-scale automation and tracking to the execution of common algorithms on standardized data, our GDAC Firehose helps to (a) facilitate innovation by allowing researchers to focus more on novel science than routine data processing, characterization and interpretation; (b) diminish needless duplication of effort at multiple research institutes; (c) democratize and disseminate TCGA science by lowering entry barriers for the global community; and (d) enhance rigor, transparency and scientific reproducibility. As a result Firehose now enjoys a central role in TCGA and has contributed results to numerous publications, portals, and research efforts in both academia and industry.