FIREBROWSE : Mining the Firehose of TCGA

Michael S. Noble[1], Katherine Huang[1], Kane Hadley[1], Noam Shoresh[1], Eila Arich-Landkof[1,2], Benjamin Alexander[1], David I. Heiman[1], and Gad Getz[1,2]

[1]Broad Institute of MIT and Harvard, Cambridge, MA ■ [2]Massachusetts General Hospital, Boston, MA

The Broad Institute GDAC Firehose

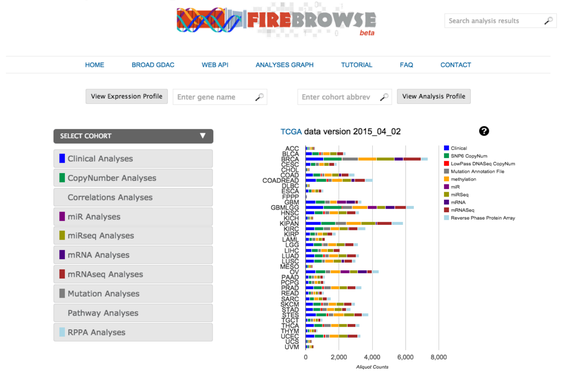
Born of the desire to systematize analyses from The Cancer Genome Atlas pilot and scale their execution to the dozens of remaining diseases to be studied, ***Firehose*** *now sits atop ~54 terabytes of TCGA data and reliably executes thousands of pipelines per month.*

- Version-stamped, standardized datasets
- Version-stamped packages of standard scientific analysis results
- Version-stamped custom runs for TCGA analysis working groups

Making the results of such available through biologist-friendly and literature-citable online reports, and en masse through ***`firehose_get`***

Simplified Portal Access

In the summer of 2014 we introduced firebrowse.org, which makes it easy to find any of the thousands of TCGA datasets or Firehose analysis result reports in just 2 clicks.



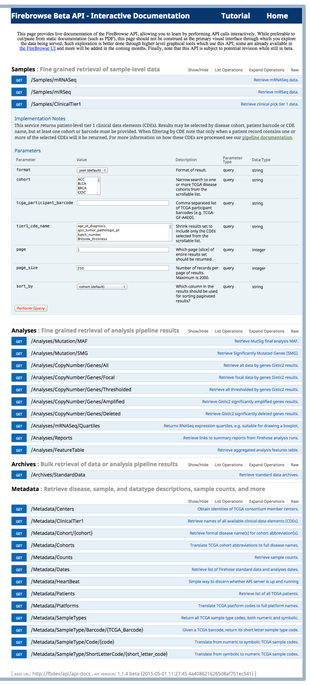
RESTful API

Powering the FireBrowse GUI is a RESTful application programming interface (API) with 27 functions and growing, fully open to the public.

Gives bulk or fine-grained access to:

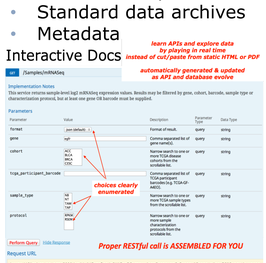
- Sample-level data
- Firehose analyses
- Standard data archives
- Metadata

Interactive Docs

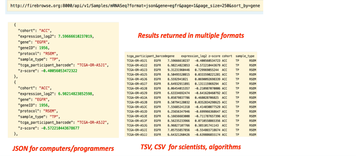


learn APIs and explore data by playing in real time instead of curl/paste from static HTML or PDF

automatically generated & updated as API and database evolve



Results returned in multiple formats



JSON for computers/programmers

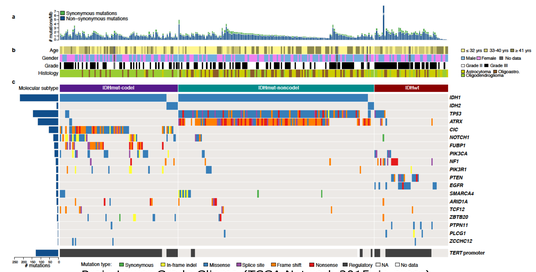
TSV, CSV for scientists, algorithms

New Features in FireBrowse

- Continuing to grow out the API and interactive documentation
  - Quartiles for mRNASeq expression
  - Patients for participant barcodes, optionally filtered by cohort
- Open source client-side wrappers for the RESTful API:
  - `fbget` - https://confluence.broadinstitute.org/display/GDAC/fbget
    - High- and low-level Python bindings
    - The `fbget` CLI, for easy access from UNIX command line
    - Automatically generated to keep synchronized with the RESTful API
  - `FireBrowseR` - https://github.com/mariodeng/FirebrowseR
    - R bindings developed by a Ph.D candidate in Germany
- Visualization tools:
  - iCoMut - interactive visualization of mutation co-occurrence
  - viewGene - a mRNASeq gene expression viewer

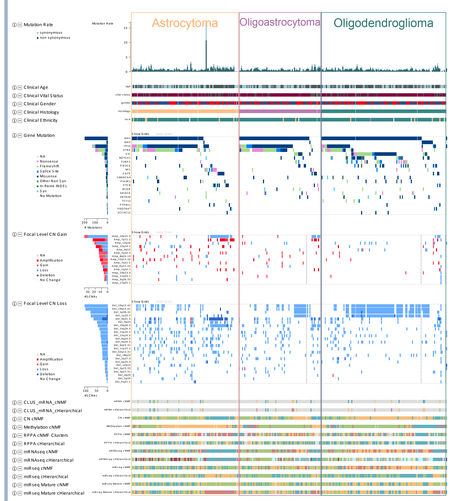
iCoMut

CoMut plots are a staple of many TCGA papers. They provide a comprehensive analysis profile in a single graphic, enabling the reader to infer relationships between co-occurring results.

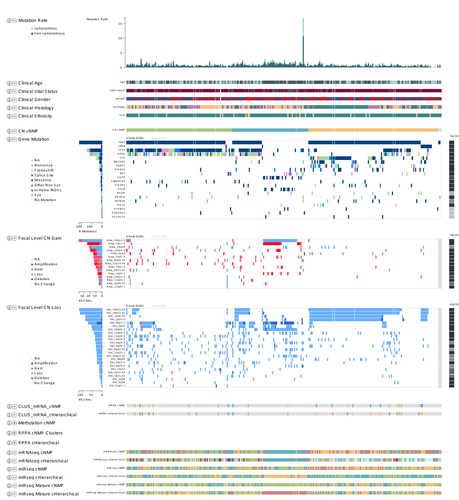


Brain Lower Grade Glioma (TCGA Network 2015, in press)

With iCoMut researchers can now play with coMut plots interactively, sorting and reordering the samples and results as they see fit; this provides a powerful synoptic tool for interpretation and exploration.



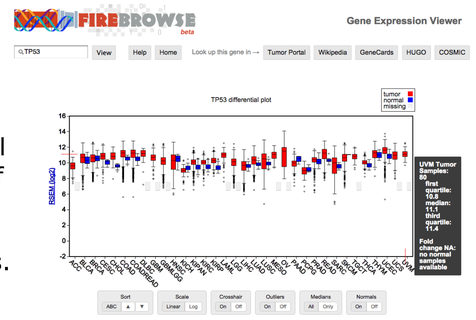
The left panel is very close to what iCoMut displays out-of-the-box: sorted first by the histology clinical parameter, then by gene (initially ordered by descending mutation count). It is quickly apparent that copy-number changes differ when IDH1/2, TP53, and ATRX mutations drop off.



In the right panel we re-sort by copy-number clustering, making it further apparent that not only is the copy-number landscape different with the lack of aforementioned mutations, there also seems to be involvement with EGFR and PTEN.

viewGene

Utilizing the RESTful API, the viewGene tool generates a boxplot of mRNASeq expression levels for a selected gene across all cohorts.



What's new in GDAC Firehose?

Raw MAFs

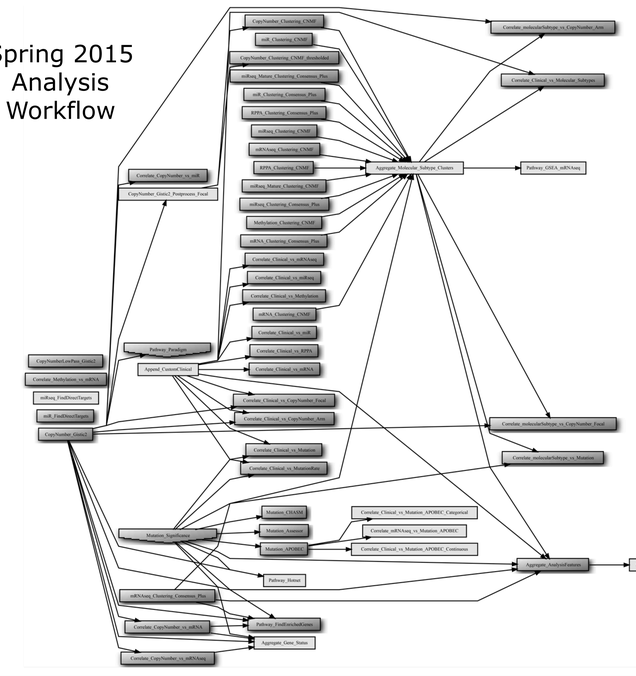
For many cancer types, mutation samples continued to be sequenced after paper publication. Previously, we only packaged and ran analyses on the mutations submitted with the paper in such cases. We are excited to announce that our latest analyses run also include post-publication samples, so far ***adding over 1500 mutation samples to our data stream.***

New Analyses

7 new analyses in our latest run, over 300 new reports:

- Correlate_Clinical_vs_Mutation_APOBEC_Categorical
- Correlate_Clinical_vs_Mutation_APOBEC_Continuous
- Correlate_mRNAseq_vs_Mutation_APOBEC
- miRseq_FindDirectTargets
- Mutation_CoOccurrence
- Pathway_GSEA_mRNAseq
- Pathway_Overlaps_MSigDB_MutSig2CV

Spring 2015 Analysis Workflow



BROAD INSTITUTE

MASSACHUSETTS GENERAL HOSPITAL

FIREHOSE Broad GDAC

The Cancer Genome Atlas