Firehose Pipeline Initiatives

David I. Heiman[1], Juok Cho[1], Daniel DiCara[1], Nils Gehlenborg[1,2], Jaegil Kim[1], Yingchun Liu[1], Gordon Saksena[1], Hailei Zhang[1], Lihua Zou[1], and Michael S. Noble[1]

[1]Broad Institute of MIT and Harvard, Cambridge, MA

[2]Center for Biomedical Informatics, Harvard Medical School, Boston, MA

Firehose AWG Runs

Purpose: The Broad Institute GDAC will gladly coordinate with TCGA analysis working groups (AWGs) to provide custom Firehose runs tailored to their specific needs. This represents an evolution of Firehose, beyond its original mission of monthly runs intended for archival storage at the DCC and wide public consumption beyond TCGA, by providing in-depth support to ongoing analysis efforts within TCGA. This provides several "real-time value-added benefits" to AWGs:

- Currency: pipelines can be run on the latest daily snapshot of data from the DCC, avoiding the time & sample lag of monthly runs
- Flexibility: additional runs can be easily performed on AWG-curated disease subtypes, and even include custom analyses
- Speed: custom AWG runs can be executed in only a few days time (excluding computationally intensive algorithms that may take >1 week to run)
- Familiarity: using the same internal Firehose machinery, and external-facing dashboards, Nozzle reports etc, already known to the community
- Easy Reproducibility: single system as starting point, everything kept, runs will soon be reference-able by DOI
- Scope: a stepping stone to open-access Firehose, that can be manipulated directly by TCGA researchers, instead of having runs curated only at the Broad

Our custom AWG data runs can also be used to define a baseline AWG data freeze. These freeze products are ideally suited for sharing data across the various centers participating in a given AWG. Furthermore, all of the output archives produced by our custom AWG runs are easily obtained with firehose_get, in the same manner as the monthly runs.

DiseaseType	AWG Run Dashboard
GBM	2013_02_17
LGG	2013_04_06 2013_02_03 2013_01_16
HNSC	2013_03_30
LUAD	2013_02_07 2012_11_15
PANCAN8	2012_08_25
SKCM	2013_01_16 2012_12_21
STAD	2013_04_17
THCA	2013_03_18 2012_10_24

TCGA Firehose AWG analyses for THCA: 2013_03_18

Maintained by TCGA GDAC Team (Broad Institute/MD Anderson Cancer Center/Harvard Medical School)

Data snapshot March 18th, leading to freeze in two weeks

Unique Tumor Sample Counts

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
THCA	500	318	432	94	500	0	409	0	411	224	323

Download run results with firehose_get

Example download command: ***firehose_get awg_thca 2013_03_18***

Analysis workflow run 5 times: On tumor primary cohort And 4 histological subtypes

Cancer Types

Table 1. Click "Browse" to view reports for a cancer type of interest. If you prefer to view reports on your own computer, you may download a ZIP archive containing all reports for a cancer type by clicking "Download".

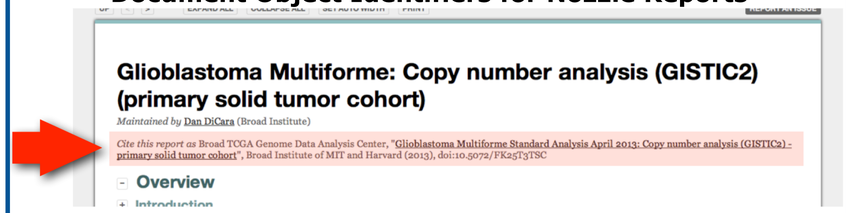
Cancer Type	Cohort	Reports	HTML	ZIP
Thyroid Adenocarcinoma	THCA-TP	23	Browse	Download
Thyroid Adenocarcinoma	THCA-classical	10	Browse	Download
Thyroid Adenocarcinoma	THCA-follicular	10	Browse	Download
Thyroid Adenocarcinoma	THCA-other	4	Browse	Download
Thyroid Adenocarcinoma	THCA-tall-cell	8	Browse	Download

225+ pipelines executed 55 analysis reports In ~2 days

Download with firehose_get Or browse from desktop

http://gdac.broadinstitute.org/runs/awg_thca__2013_03_18/

Document Object Identifiers for Nozzle Reports



All future Firehose Nozzle reports published by the Broad Institute GDAC will be assigned a permanent Document Object Identifier (DOI). By assigning DOIs to standard analysis run and Analysis Working Group run reports, we request that scientists who use Firehose results in their studies directly cite the source of their data.

Document Object Identifiers (DOIs) are identifiers used commonly to reference publications in scientific literature. DOIs are resolved into URLs through a central service such as http://dx.doi.org. Unlike URLs, DOIs are permanent and can be associated with modifiable meta-data such as authors, publisher, and publication date. Journals and other scientific publishers prefer DOIs in archival literature citations due to their permanency and flexibility.

The DOIs used for Nozzle Reports are provided and managed by the California Digital Libraries EZID service (http://www.cdlib.org/services/uc3/ezid), a member of the DataCite organization (http://www.datacite.org). DataCite is one of eight DOI registration agencies, and focuses on working with organizations that generate and hold research data.

The Nozzle R package* has been extended to support the insertion of a DOI and citation information. The updated source code is available at http://www.github.com/parklab/Nozzle and will be part of the upcoming Nozzle 1.1 release.

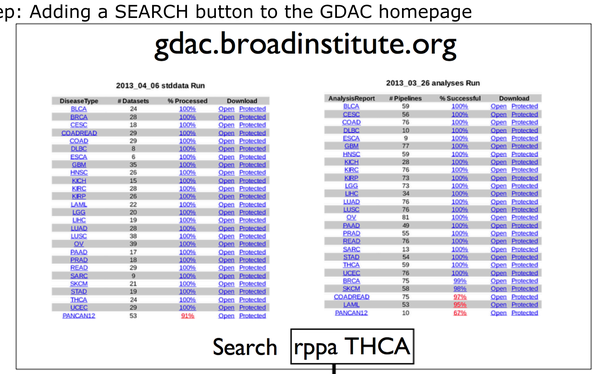
* N Gehlenborg, MS Noble, G Getz, L Chin and PJ Park, "Nozzle: a report generation toolkit for data analysis pipelines", *Bioinformatics* **29**:1089-1091 (2013).

Additional Web Search Capabilities

Broad GDAC executes over 3,000 pipelines monthly, and publishes more than 700 online reports.

This copious content is automatically indexed by an internal Broad Google appliance, the external Google web crawler, and is visible in IGV (cBio, GenomeBrowser, …)

The Next Step: Adding a SEARCH button to the GDAC homepage



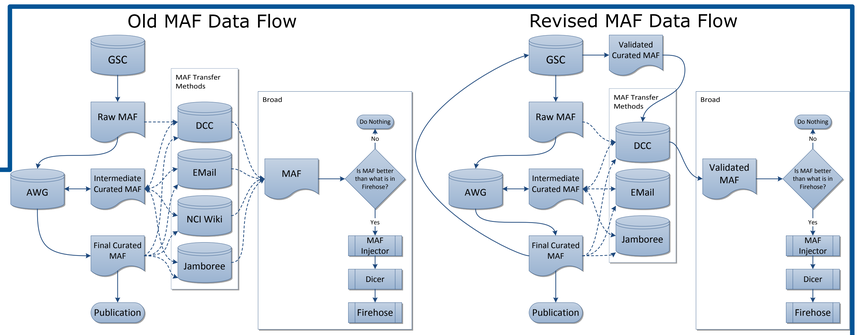
MAF Curation: Improving Consistency & Transparency via Policy & Automation

Historically, tracking MAFs has been difficult. There was no guarantee that a final AWG curated MAF would even make it to the DCC, let alone into Firehose. In order to keep the MAFs ingested into Firehose up-to-date, the maintainer had to track multiple AWG email lists, the AWGs' private wiki pages, the Jamboree site, and the DCC. Even then, it often came down to receiving an email from an analyst asking why the latest MAF wasn't available in Firehose.

Often these MAFs did not come from the DCC, requiring formatting cleanup and possible removal of germline data. Sometimes required data fields would be empty. Unvalidated MAFs and lack of a centralized repository led to inconsistent data.

To remedy this inconsistency, we now enforce the revised MAF data flow (pictured on the right). We now require that any MAF we ingest into Firehose MUST exist outside of it in a publically accessible fashion.

This enforcement has enabled us to better streamline our overall process. We have developed tools to automatically parse the MAFs that we mirror from the DCC and the MAFs we are ingesting into Firehose, summarize the information, and generate Confluence pages with these summaries laid out in an easy to read manner (partial examples on right).



MAFs Available from the DCC as of 25 April 2013

Tumor Type	Center	Tumor Samples	Normal Samples	Mutations	Columns	Archive Version	File Name	Size	Last Modified	md5sum
BLCA	broad.mit.edu	136	145	44875	90	0.2.0	PR_TCGA_BLCA_PAIR_Capture_All_Pairs_QCPASS_v2.aggregated.capture.tcga.uuid.somatic.maf	38.4MB	27 Mar 2013	dd846e7240ff5ba72397412fff3b3f3c
	broad.mit.edu	28	28	7557	86	1.3.0	BLCA-28-original.aggregated.tcga.somatic.maf	6.5MB	09 Jan 2013	816f494b9caf23b944249651a7d2cded
BRCA	genome.wustl.edu	107	109	7746	54	2.1.0	genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.2.0.0.maf	4.0MB	09 Oct 2012	1d662b8319e3a2672db293b15252ed4f
	genome.wustl.edu	510	546	29839	55	3.3.0	genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.3.2.0.somatic.maf	17.8MB	11 Oct 2012	dbc316c2f817dfc4284df434a5d152c8
	genome.wustl.edu	776	801	47243	55	5.2.0	genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.5.1.0.somatic.maf	24.9MB	09 Oct 2012	ffa233efe6085306bff0ffdef0d0736e
CESC	broad.mit.edu	39	39	10020	90	1.4.0	PR_TCGA_CESC_PAIR_Capture_All_Pairs.aggregated.capture.tcga.uuid.somatic.maf	8.1MB	12 Nov 2012	1b13e812e746b9149b15090b48d2de1e
	broad.mit.edu	40	40	11509	90	0.2.0	PR_TCGA_CESC_PAIR_Capture_All_Pairs_QCPASS_v2.aggregated.capture.tcga.uuid.somatic.maf	9.3MB	27 Mar 2013	0422e7b9be78aa6d7a649f11090f120c
COAD	hgsc.bcm.edu	220	221	114594	47	1.4.0	hgsc.bcm.edu_COAD.IlluminaGA_DNASeq.1.somatic.maf	42.1MB	18 Apr 2013	adb65259de592b979d9134dcceb6a8bd
	hgsc.bcm.edu	53	53	10986	41	1.5.0	hgsc.bcm.edu_COAD.SOLiD_DNASeq.1.somatic.maf	3.4MB	25 Mar 2013	7e97bc803951045576ff8c5479d84be4
GBM	broad.mit.edu	291	291	22166	79	1.1.0	gbm_liftover.aggregated.capture.tcga.uuid.somatic.maf	19.2MB	27 Feb 2013	1ee5e9b937ef29b1b3ccc0f6b454f886

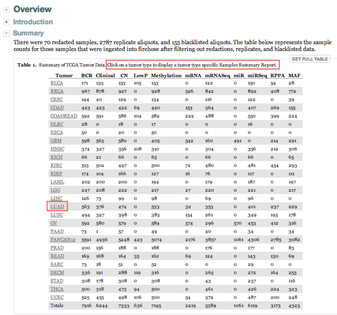
MAFs Ingested into Broad GDAC Firehose as of 26 April 2013

Tumor Type	Center	Tumor Samples	Normal Samples	Mutations	Columns	File Name	Size	Last Modified	md5sum
BLCA	broad.mit.edu	28	28	7557	86	BLCA-28-original.aggregated.tcga.somatic.maf.txt	6.5MB	20 Feb 2013	816f494b9caf23b944249651a7d2cded
BRCA	genome.wustl.edu	776	801	47243	55	genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.5.1.0.somatic.maf.txt	24.9MB	22 Feb 2013	ffa233efe6085306bff0ffdef0d0736e
CESC	broad.mit.edu	39	39	10020	90	PR_TCGA_CESC_PAIR_Capture_All_Pairs.aggregated.capture.tcga.uuid.somatic.maf.txt	8.1MB	20 Feb 2013	1b13e812e746b9149b15090b48d2de1e
COAD	bcm.edu	53	53	11578	39	hgsc.bcm.edu-COAD-from-BCM-Firehose-CRC-08nov11.maf.txt	2.8MB	18 Dec 2012	6065b55a82518412be125d55a6307179
	broad.mit.edu	102	102	50982	39	hgsc.bcm.edu_COAD.IlluminaGA_DNASeq.1.broad.mit.edu.maf.txt	12.8MB	23 Jul 2012	b55060ac2c5f4d203faec91622a2b427
GBM	broad.mit.edu	291	291	22166	79	gbm_liftover.aggregated.capture.tcga.uuid.somatic.maf.txt	19.2MB	01 Mar 2013	1ee5e9b937ef29b1b3ccc0f6b454f886
HNSC	broad.mit.edu	306	364	74008	89	PR_TCGA_HNSC_PAIR_Capture_TP-NT_TP-NB.aggregated.capture.tcga.uuid.somatic.maf.txt	61.7MB	17 Dec 2012	90c3511c08006f651117e369e5a3c243
KICH	bcm.edu	65	65	1788	48	hgsc.bcm.edu_KICH.IlluminaGA_DNASeq.1.somatic.maf.txt	669.4KB	28 Mar 2013	591c5b7e699ea7761daa932e409ebda3
KIRC	bcm.edu	176	176	8768	32	BCM-BI-Renal-v1.3.hgsc.bcm.edu.maf.txt	1.8MB	23 Nov 2011	ed030b92cf604abfb3fb4a7d9186d235
	broad.mit.edu	297	297	95839	84	BI_and_BCM_1.4.aggregated.tcga.broad.mit.edu.maf.txt	58.9MB	30 Jul 2012	2ab7c89d11a07761805 0fb4e6f5390ca

Samples Summary Reports

Now reflect significantly higher-resolution metadata per disease cohort:

- In addition to the combined list of all excluded samples (across all diseases) the report now lists them for each disease cohort as well.
- E.g. LUAD: 1 sample was redacted, 126 were filtered as replicates, and 6 were blacklisted.



- Data resolution down to the aliquot is now available for both included and excluded samples:
  - sequencing platform
  - submitting center
  - data level
  - protocol
  - sample type (e.g. primary tumor, recurrent tumor, normal blood, etc.)