Firehose in the TCGA

International Cancer Genome Consortium
Bioinformatics Analysis Working Group Telecon
June 28, 2012

Michael S. Noble
The Broad Institute of MIT & Harvard

BROAD INSTITUTE

THE CANCER GENOME ATLAS

FIREHOSE
Broad GDAC

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, now sits atop 14+ terabytes of TCGA data and reliably executes more than 1000 pipelines per month.

Because The Bad Old Days ...

Of solitary, manual experimentation ...

```
% create a folder
% download data.from.some.where
% run_your_computational_analysis
```

Then do it again Nov 13, 17, ...
Then forget ... and search, search, search
Then repeat ALL for 19 more tumors
GBM, LUNG, AML, ...

Then multiply by 5, 10 ... researchers at your site

Doesn't Scale to TCGA

New RPPA datatype
+2087 protein samples

+917
Methylation

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	89	65	58	78	0	32	0	54	0	28
BRCA	859	857	833	858	529	751	0	781	408	507
CESC	110	25	68	0	0	0	0	8	0	36
COADREAD	590	590	575	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	563	546	287	542	0	491	0	214	276
HNSC	294	255	165	292	0	103	0	89	0	0
KIRC	502	502	490	500	72	469	0	463	454	327
KIRP	135	84	75	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	144	140	143	0	27	0	0	30	0	0
LIHC	84	55	58	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	372	274	266	347	32	106	0	95	0	147
LUSC	290	272	282	282	154	220	0	202	0	178
OV	592	580	564	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	153	0	100	153	0	0	0	63	0	0
SKCM	253	0	219	240	0	0	0	0	0	0
STAD	162	150	132	133	0	57	0	123	0	133
THCA	274	73	228	230	0	0	0	45	0	0
UCEC	462	425	430	451	54	266	0	359	200	239
PANCANCER	6239	5110	5246	5325	2218	2297	1055	2844	2087	2610

April 2012 samples in Firehose (with differentials)

So Firehose Automatically Generates

1. Standardized datasets

   Aggregated, version-stamped

   Analysis-ready format / semantics

   Twice per month

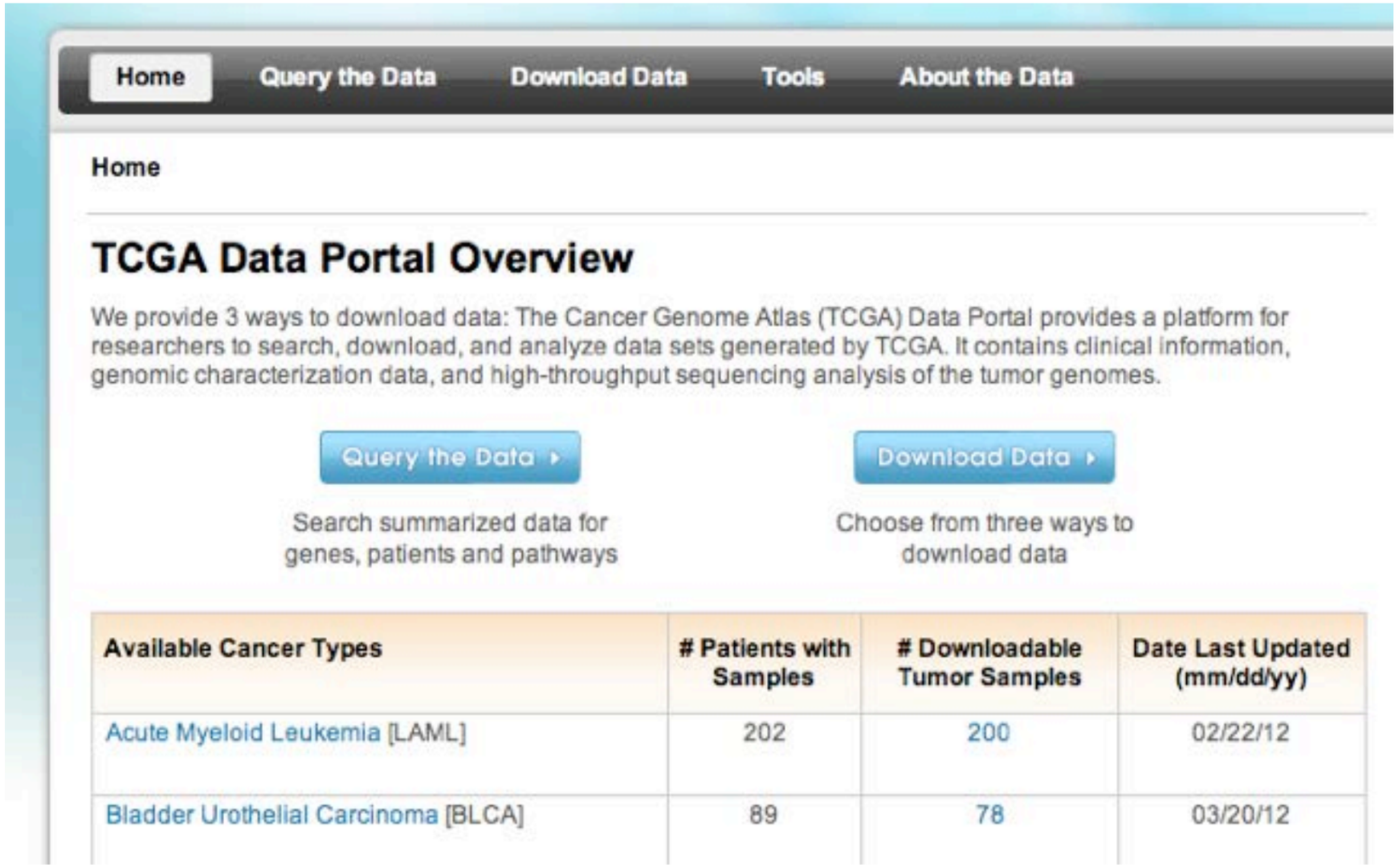
2. Standardized analyses upon them

   For vetted algorithms: GISTIC, MutSig, CNMF, ...

   Companioned with biologist-friendly reports

   Once per month

But why Firehose ...



... when TCGA data portal already exists?

Because TCGA data portal is more “raw” ...

No aggregate/versioning: hundreds of micro-versioned files

Inconsistencies across data submitted by multiple centers

So, how to use portal data directly in my research?

Are they homogeneous?

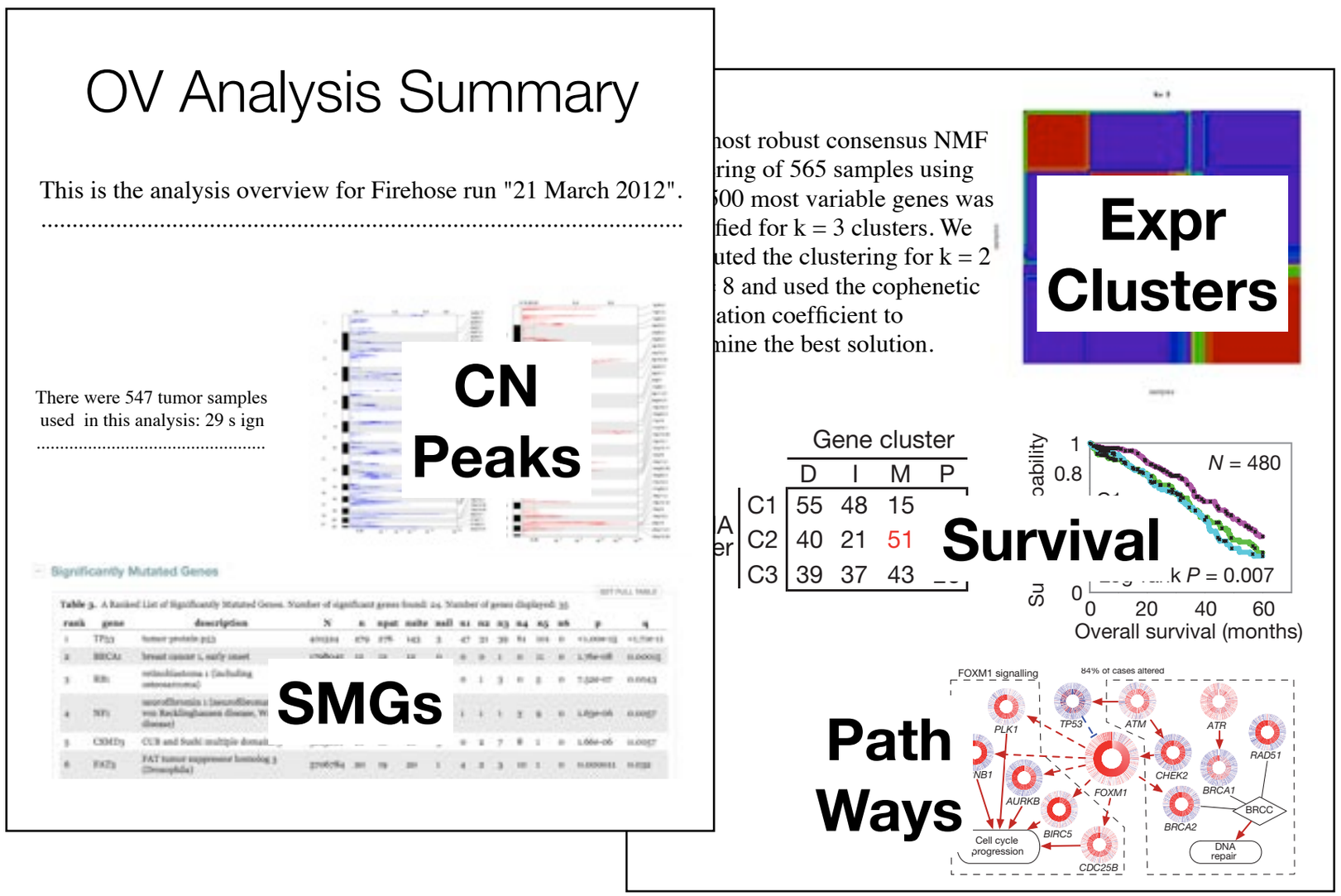
Or systematically prepared?

To be ready to load in my R or MatLab script?

We had to do this, so would you

... and does not encompass higher-level analyses

What if I just want to view GBM gistic (CN) peaks?
Or look at a methylation cluster?
Or see how expression in OV correlates to survival?



You might otherwise need to ...

Spend weeks obtaining protected data credentials

Or becoming a TCGA data guru

And still more time, mastering the analytics

Complexity & volume preclude

this approach for many teams/individuals

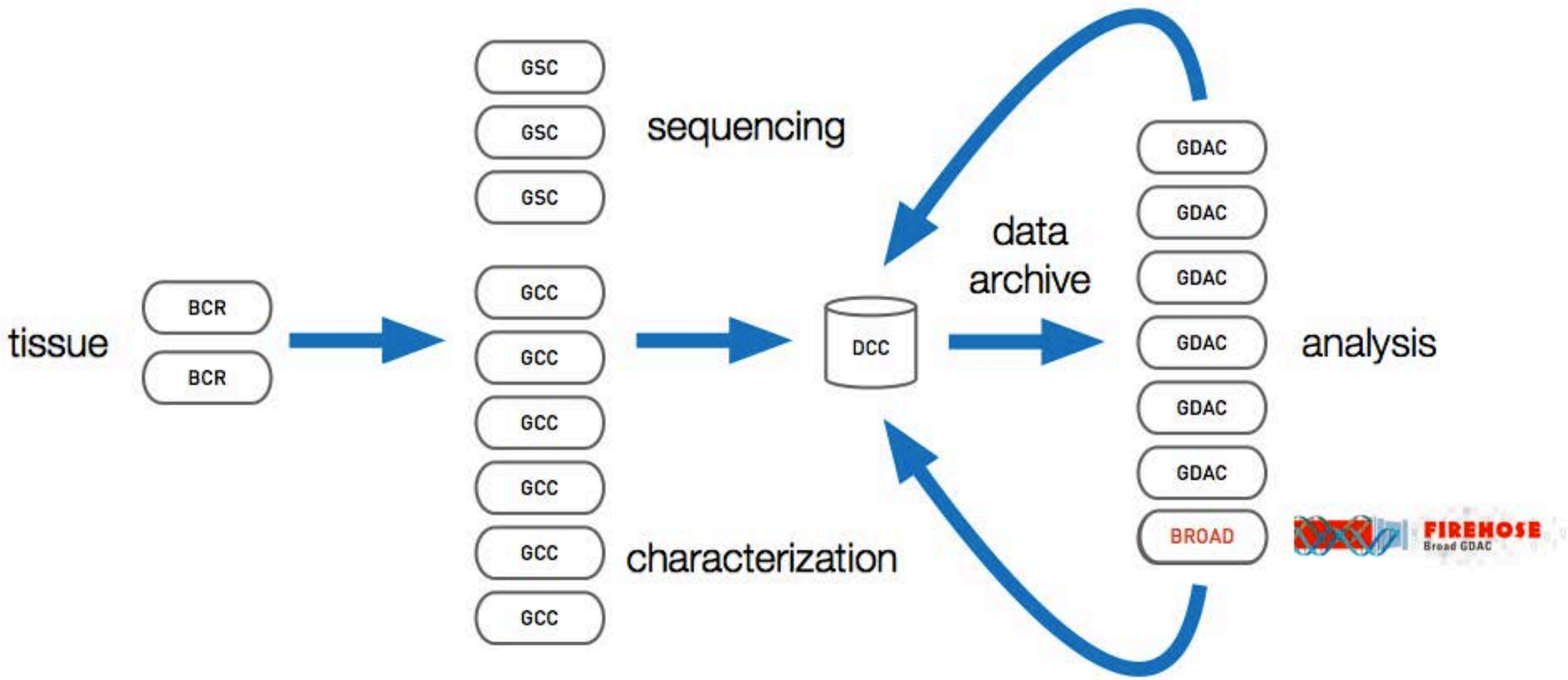
FIREHOSE
Broad GDAC

Data Standardization 1
+
Common Algorithms 2

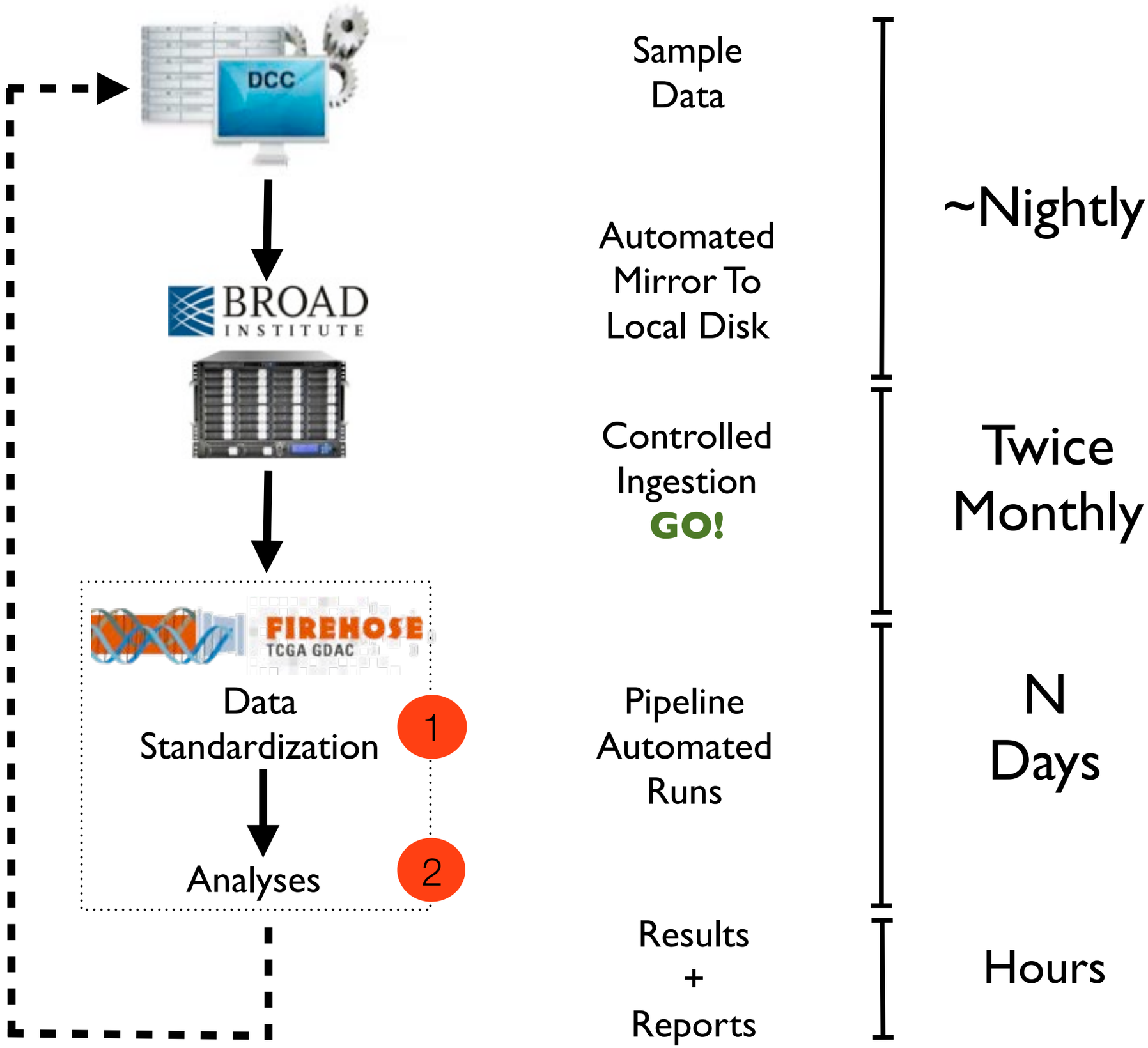
GDAC = A Data Factory

Value Added : Automatic, turnkey service for ENTIRE cancer community

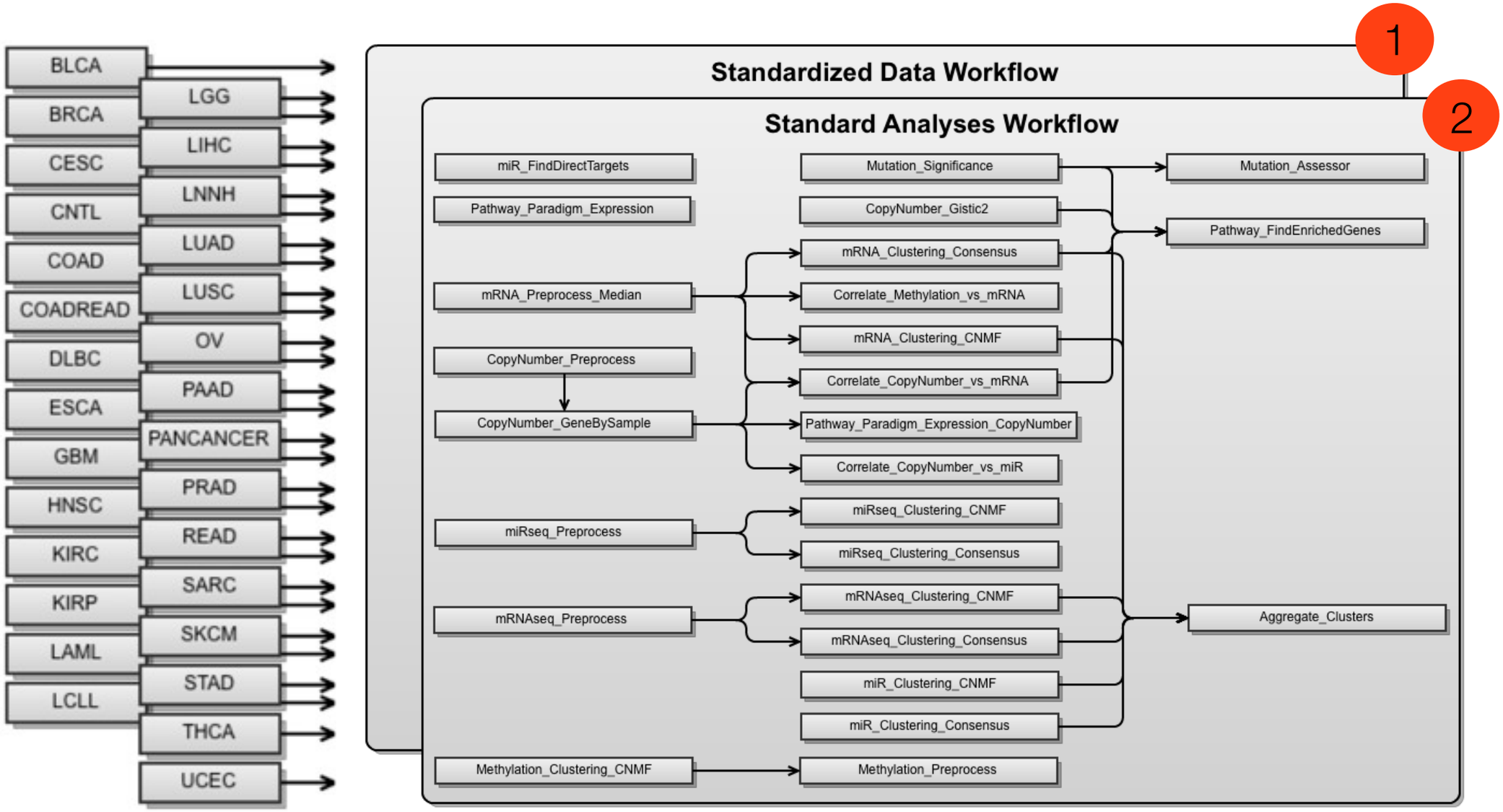
WHERE FIREHOSE LIVES



Flow of Data & Runs



KiloPipeline Per Month



Standardized Data:	273 platforms (across 23 tumorsets) x 2/month	= 546
Standardized Analyses:	(Up to) 33 analyses x 23 tumorsets / month	> 500

Ok, so far so good, but then where are ...

Standardized, ***analysis-ready*** TCGA data?

And ***standardized analyses*** upon them?

FIREHOSE
Broad GDAC

http://gdac.broadinstitute.org

Welcome to the online home of the Broad Institute's Genome Data Analysis Center (GDAC). On behalf of The Cancer Genome Atlas (TCGA), we've designed and operate scientific data and analysis pipelines which pump terabyte-scale genomic datasets through scores of quantitative algorithms, in the hope of accelerating the understanding of cancer. See the dashboards below for details of the latest monthly runs, or this presentation for more background information. Note that downloading data from our site constitutes agreement to this data usage policy.

2012_05_25 stddata Run

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	20	100%	Open	Protected
DLBC	1	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC			Open	Protected
KIRC			Open	Protected
KIRP			Open	Protected
LGG			Open	Protected
LIHC			Open	Protected
LNNH			Open	Protected
LUAD			Open	Protected
LUSC			Open	Protected
OV	29	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	10	100%	Open	Protected
SKCM	6	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	13	100%	Open	Protected
UCEC	22	100%	Open	Protected
LAML	10	91%	Open	Protected
PANCANCER	40	83%	Open	Protected

Data Dashboard

2012_05_25 analyses Run

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	14	100%	Open	Protected
BRCA	27	100%	Open	Protected
COADREAD	27	100%	Open	Protected
GBM	26	100%	Open	Protected
HNSC	11	100%	Open	Protected
KIRC	27	100%	Open	Protected
LAML			pen	Protected
LGG			pen	Protected
LIHC			pen	Protected
LUSC			pen	Protected
OV			pen	Protected
PAAD			pen	Protected
PRAD			pen	Protected
SKCM			pen	Protected
STAD	16	100%	Open	Protected
THCA	8	100%	Open	Protected
UCEC	27	100%	Open	Protected
LUAD	21	95%	Open	Protected
KIRP	15	94%	Open	Protected
CESC	4	75%	Open	Protected
PANCANCER	6	35%	Open	Protected

Analysis Dashboard

View analysis reports or click on dashboards above or download with firehose_get.

1 Standardized Data Dashboard

2012_05_25 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	20	100%	Open	Protected
DLBC	1	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	22	100%	Open	Protected
KIRP	16	100%	Open	Protected
LGG	11	100%	Open	Protected
LIHC	12	100%	Open	Protected
LNNH	1	100%	Open	Protected
LUAD	20	100%	Open	Protected
LUSC	29	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	10	100%	Open	Protected
SKCM	6	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	13	100%	Open	Protected
UCEC	22	100%	Open	Protected
LAML	10	91%	Open	Protected
PANCANCER	40	83%	Open	Protected

View: Rationale Release notes FAQ Download: firehose_get

Fast/Simple
Overview Of
Analysis-ready data

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	100	66	35	78	0	56	0	54	0	28
BRCA	871	857	781	858	529	777	0	781	408	507
CESC	110	26	36	0	0	0	0	8	0	36
COADREAD	590	590	564	584	224	83	0	255	399	224
DLBC	27	0	0	0	0	0	0	0	0	0
GBM	595	564	537	287	542	0	491	0	214	276
HNSC	312	283	165	292	0	263	0	89	0	0
KIRC	502	502	489	500	72	469	0	463	454	327
KIRP	135	95	43	117	16	14	0	16	0	0
LAML	202	200	0	192	0	179	0	187	0	199
LGG	181	144	80	0	27	0	0	30	0	0
LIHC	84	55	53	0	0	17	0	28	0	0
LNNH	2	0	0	0	0	0	0	0	0	0
LUAD	409	292	205	347	32	129	0	95	0	229
LUSC	326	279	211	282	154	223	0	202	0	178
OV	592	580	547	551	568	0	564	46	412	316
PAAD	48	0	14	30	0	0	0	0	0	0
PRAD	172	0	82	153	0	0	0	63	0	83
SARC	21	0	0	0	0	0	0	0	0	0
SKCM	253	0	0	240	0	0	0	0	0	0
STAD	162	155	132	133	0	57	0	123	0	133
THCA	300	158	85	230	0	3	0	45	0	0
UCEC	462	425	363	451	54	266	0	359	200	248
PANCANCER	6456	5271	4422	5325	2218	2536	1055	2844	2087	2784

1 Standardized Data Dashboard

2012_05_25 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected

Or drill down for detailed info

e.g. showing 2 methylation platforms, and originating center (Johns Hopkins)

Broad GDAC Standard Data Status
stddata__2012_05_25 Run for Tumor Type: BRCA

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BRCA	871	857	781	858	529	777	0	781	408	507

Note that the links below require Broad internal Firehose login credentials.

	Pipeline Dataset	Not Available	Available	InProcess	Successful	Unsuccessful
1	Clinical_Pick_Tier1	0	0	0	1	0
2	Merge_Clinical	0	0	0	1	0
3	Merge_cna__cgh_1x1m_g4447a__mskcc_org__Level_2__bioassay_data_transformation__data	1	0	0	0	0
4	Merge_cna__cgh_1x1m_g4447a__mskcc_org__Level_3__segmentation_data_computation__seg	1	0	0	0	0
5	Merge_cna__hg_cgh_244a__hms_harvard_edu__Level_2__lowess_global_normalization__data	1	0	0	0	0
6	Merge_cna__hg_cgh_244a__hms_harvard_edu__Level_3__segmentation__seg	1	0	0	0	0
7	Merge_cna__hg_cgh_244a__mskcc_org__Level_2__bioassay_data_transformation__data	1	0	0	0	0
8	Merge_cna__hg_cgh_244a__mskcc_org__Level_3__segmentation_data_computation__seg	1	0	0	0	0
9	Merge_cna__hg_cgh_415k_g4124a__hms_harvard_edu__Level_2__lowess_global_normalization__data	1	0	0	0	0
10	Merge_cna__hg_cgh_415k_g4124a__hms_harvard_edu__Level_3__segmentation__seg	1	0	0	0	0
11	Merge_exon__huex_1_0_st_v2__lbl_gov__Level_2__quantile_normalization_exon__data	1	0	0	0	0
12	Merge_exon__huex_1_0_st_v2__lbl_gov__Level_3__quantile_normalization_gene__data	1	0	0	0	0
13	Merge_exon__huex_1_0_st_v2__lbl_gov__Level_3__segmented_as_firma__data	1	0	0	0	0
14	Merge_methylation__humanmethylation27__jhu_usc_edu__Level_2__within_bioassay_data_set_function__data	1	0	0	0	0
15	Merge_methylation__humanmethylation27__jhu_usc_edu__Level_3__within_bioassay_data_set_function__data	0	0	0	1	0
16	Merge_methylation__humanmethylation450__jhu_usc_edu__Level_3__within_bioassay_data_set_function__data	0	0	0	1	0
17	Merge_methylation__illuminadnamethylation_oma003_cpi__jhu_usc_edu__Level_2__within_bioassay_data_set_function__data	1	0	0	0	0
18	Merge_mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3__isoform_expression__data	1	0	0	0	0
19	Merge_mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3__mirna_expression__data	1	0	0	0	0
20	Merge_mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3__miR_gene_expression__data	0	0	0	1	0
21	Merge_mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3__miR_isoform_expression__data	0	0	0	1	0
22	Merge_mirnaseq__illuminahiseq_mirnaseq__bcgsc_ca__Level_3__miR_gene_expression__data	0	0	0	1	0
23	Merge_mirnaseq__illuminahiseq_mirnaseq__bcgsc_ca__Level_3__miR_isoform_expression__data	0	0	0	1	0

1 Standardized Data Dashboard

2012_05_25 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	20	100%	Open	Protected
DLBC	1	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	22	100%	Open	Protected
KIRP	16	100%	Open	Protected
LGG	11	100%	Open	Protected
LIHC	12	100%	Open	Protected
LNNH	1	100%	Open	Protected
LUAD	20	100%	Open	Protected
LUSC	29	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	10	100%	Open	Protected
SKCM	6	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	13	100%	Open	Protected
UCEC	22	100%	Open	Protected
LAML	10	91%	Open	Protected
PANCANCER	40	83%	Open	Protected

View: Rationale Release notes FAQ Download: firehose_get

With Redactions Provenance

EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

2012_05_25 Redactions Report

Overview

Introduction

For TCGA data, redaction is the removal of cases from the data prior to publication or release. Redacted cases are generally rare, but cases must be redacted when the TSS/BCR subject link is incorrect ("unknown patient identity"), or in the case of genotype mismatch, completely wrong cancer, or completely wrong organ/tissue. Redaction occurs regardless of a case's analyte characterization or DCC data deposition status.

Summary

There were 43 redactions.

Results

GET FULL TABLE

Table 1.

Barcode	UUID	Date	Type	Notes
TCGA-01-0629	(none)	09/02/2010	GBM	[intgen.org]: Case was of non-ovarian origin
TCGA-01-0638	(none)	09/02/2010	OV	[intgen.org]: Case was of non-ovarian origin
TCGA-02-0002	(none)	09/15/2010	GBM	[intgen.org]: Genotype mismatch
TCGA-02-0117	(none)	09/15/2010	GBM	[intgen.org]: Genotype mismatch
TCGA-02-2488	741c2eb3-db33-4a6c-9878-2587aa375134	10/27/2011	GBM	Note: Scheduled for shipment in B38 but was withdrawn due to SSTR mismatch. Case was uploaded, withdrawal was initiated after upload. Worked with DCC to remove data.
TCGA-06-0748	c8f21beb-e3ca-4e5e-a386-a486776cfe88	10/27/2011	GBM	Note: Scheduled for shipment in B8 but was withdrawn due to SSTR mismatch. Case was uploaded, withdrawal was initiated after upload. Worked with DCC to remove data.
TCGA-08-0384	(none)	11/22/2010	GBM	IGC new redactions as of 11/10/2010
TCGA-13-1479	(none)	09/02/2010	OV	[intgen.org]: Case was of non-ovarian origin
TCGA-14-0784	fc52a226-9306-45d4-b608-4bdda831ad01	09/12/2011	GBM	[intgen.org]: Genotype mismatch
TCGA-14-1036	f45a2391-4c79-4269-80dd-12a914d6c4b5	10/27/2011	GBM	Note: Scheduled for shipment in B38 but was withdrawn due to SSTR mismatch. Case was uploaded, withdrawal was initiated after upload. Worked with DCC to remove data

1 Standardized Data Dashboard

2012_05_25 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	20	100%	Open	Protected
DLBC	1	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	22	100%	Open	Protected
KIRP	16	100%	Open	Protected
LGG	11	100%	Open	Protected
LIHC	12	100%	Open	Protected
LNNH	1	100%	Open	Protected
LUAD	20	100%	Open	Protected
LUSC	29	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	10	100%	Open	Protected
SKCM	6	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	13	100%	Open	Protected
UCEC	22	100%	Open	Protected
LAML	10	91%	Open	Protected
PANCANCER	40	83%	Open	Protected

View: Rationale Release notes FAQ Download: firehose_get

Browse on site

Download interactively by mouse click

Or programmatically with ***firehose_get*** command line tool

1 Standardized Data Dashboard

2012_05_25 stddata Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	15	100%	Open	Protected
BRCA	22	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	20	100%	Open	Protected
DLBC	1	100%	Open	Protected
GBM	27	100%	Open	Protected
HNSC	15	100%	Open	Protected
KIRC	22	100%	Open	Protected
KIRP	16	100%	Open	Protected
LGG	11	100%	Open	Protected
LIHC	12	100%	Open	Protected
LNNH	1	100%	Open	Protected
LUAD	20	100%	Open	Protected
LUSC	29	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	10	100%	Open	Protected
SKCM	6	100%	Open	Protected
STAD	18	100%	Open	Protected
THCA	13	100%	Open	Protected
UCEC	22	100%	Open	Protected
LAML	10	91%	Open	Protected
PANCANCER	40	83%	Open	Protected

View: Rationale Release notes FAQ Download: firehose_get

With Supporting Documentation

Release Notes
Rationale
Frequently Asked Questions

The Broad GDAC standardized data packages represent a frozen snapshot of all TCGA analysis data at a given time:

- ***Cast in a form amenable to immediate algorithmic analysis*** (no additional data preparation required)
- Which provides a ***consistent point of reference*** for analysis and citation by marker papers and users of TCGA data
- Towards a ***formal definition*** of what constitutes a given tumor dataset
- While ***minimizing redundant effort*** across centers and groups to download & prepare data for further analysis
- And ***enhancing provenance and reproducibility***

Ok, that covers the data, but ...

What about that GISTIC peak?

Or methylation & expression cluster?

2 Analyses Dashboard

2012_05_25 analyses Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	14	100%	Open	Protected
BRCA	27	100%	Open	Protected
COADREAD	27	100%	Open	Protected
GBM	26	100%	Open	Protected
HNSC	11	100%	Open	Protected
KIRC	27	100%	Open	Protected
LAML	11	100%	Open	Protected
LGG	14	100%	Open	Protected
LIHC	8	100%	Open	Protected
LUSC	25	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	2	100%	Open	Protected
PRAD	8	100%	Open	Protected
SKCM	2	100%	Open	Protected
STAD	16	100%	Open	Protected
THCA	8	100%	Open	Protected
UCEC	27	100%	Open	Protected
LUAD	21	95%	Open	Protected
KIRP	15	94%	Open	Protected
CESC	4	75%	Open	Protected
PANCANCER	6	35%	Open	Protected

View: Analysis reports Release notes FAQ Download: firehose_get

Similar Layout to Data Dashboard

Broad GDAC Analyses Status
analyses__2012_05_25__brca_00 Run for Tumor Type: BRCA

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BRCA	871	857	833	858	529	777	0	781	408	507

Note that the links below require Broad internal Firehose login credentials.

	Pipeline	NotRunnable	Runnable	InProcess	Successful	Unsuccessful
1	Aggregate_Clusters	0	0	0	1	0
2	CopyNumber_GeneBySample	0	0	0	1	0
3	CopyNumber_Gistic2	0	0	0	1	0
4	Correlate_Clinical_vs_miR	1	0	0	0	0
5	Correlate_Clinical_vs_Molecular_Signatures	0	0	0	1	0
6	Correlate_Clinical_vs_mRNA	0	0	0	1	0
7	Correlate_Clinical_vs_Mutation	0	0	0	1	0
8	Correlate_CopyNumber_vs_miR	1	0	0	0	0
9	Correlate_CopyNumber_vs_mRNA	0	0	0	1	0
10	Correlate_CopyNumber_vs_mRNAseq	0	0	0	1	0
11	Correlate_Methylation_vs_mRNA	0	0	0	1	0
12	Methylation_Clustering_CNMF	0	0	0	1	0
13	Methylation_Preprocess	0	0	0	1	0
14	miRseq_Clustering_CNMF	0	0	0	1	0
15	miRseq_Clustering_Consensus	0	0	0	1	0
16	miRseq_Preprocess	0	0	0	1	0
17	miR_Clustering_CNMF	1	0	0	0	0
18	miR_Clustering_Consensus	1	0	0	0	0
19	miR_FindDirectTargets	1	0	0	0	0
20	miR_Preprocess	1	0	0	0	0
21	mRNAseq_Clustering_CNMF	0	0	0	1	0
22	mRNAseq_Clustering_Consensus	0	0	0	1	0
23	mRNAseq_Preprocess	0	0	0	1	0
24	mRNA_Clustering_CNMF	0	0	0	1	0
25	mRNA_Clustering_Consensus	0	0	0	1	0
26	mRNA_Preprocess_Median	0	0	0	1	0
27	Mutation_Assessor	0	0	0	1	0
28	Mutation_Significance	0	0	0	1	0
29	Pathway_FindEnrichedGenes	0	0	0	1	0
30	Pathway_Paradigm_Expression	0	0	0	1	0
31	Pathway_Paradigm_Expression_CopyNumber	0	0	0	1	0
32	RPPA_Clustering_CNMF	0	0	0	1	0
33	RPPA_Clustering_Consensus	0	0	0	1	0
	Total	6	0	0	27	0

Supplemented with analysis pipeline status
(showing 27 of 33 pipelines ran for breast cancer)

Accompanied by Biologist-Friendly Reports

2012_05_25 analyses Run

Tables of Ingested Data: HTML PNG TSV Redactions: Report

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	14	100%	Open	Protected
BRCA	27	100%	Open	Protected
COADREAD	27	100%	Open	Protected
GBM	26	100%	Open	Protected
HNSC	11	100%	Open	Protected
KIRC	27	100%	Open	Protected
LAML	11	100%	Open	Protected
LGG	14	100%	Open	Protected
LIHC	8	100%	Open	Protected
LUSC	25	100%	Open	Protected
OV	29	100%	Open	Protected
PAAD	2	100%	Open	Protected
PRAD	8	100%	Open	Protected
SKCM	2	100%	Open	Protected
STAD	16	100%	Open	Protected
THCA	8	100%	Open	Protected
UCEC	27	100%	Open	Protected
LUAD	21	95%	Open	Protected
KIRP	15	94%	Open	Protected
CESC	4	75%	Open	Protected
PANCANCER	6	35%	Open	Protected

View: Analysis reports Release notes FAQ Download: firehose_get

UP < > EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by TCGA GDAC Team (Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

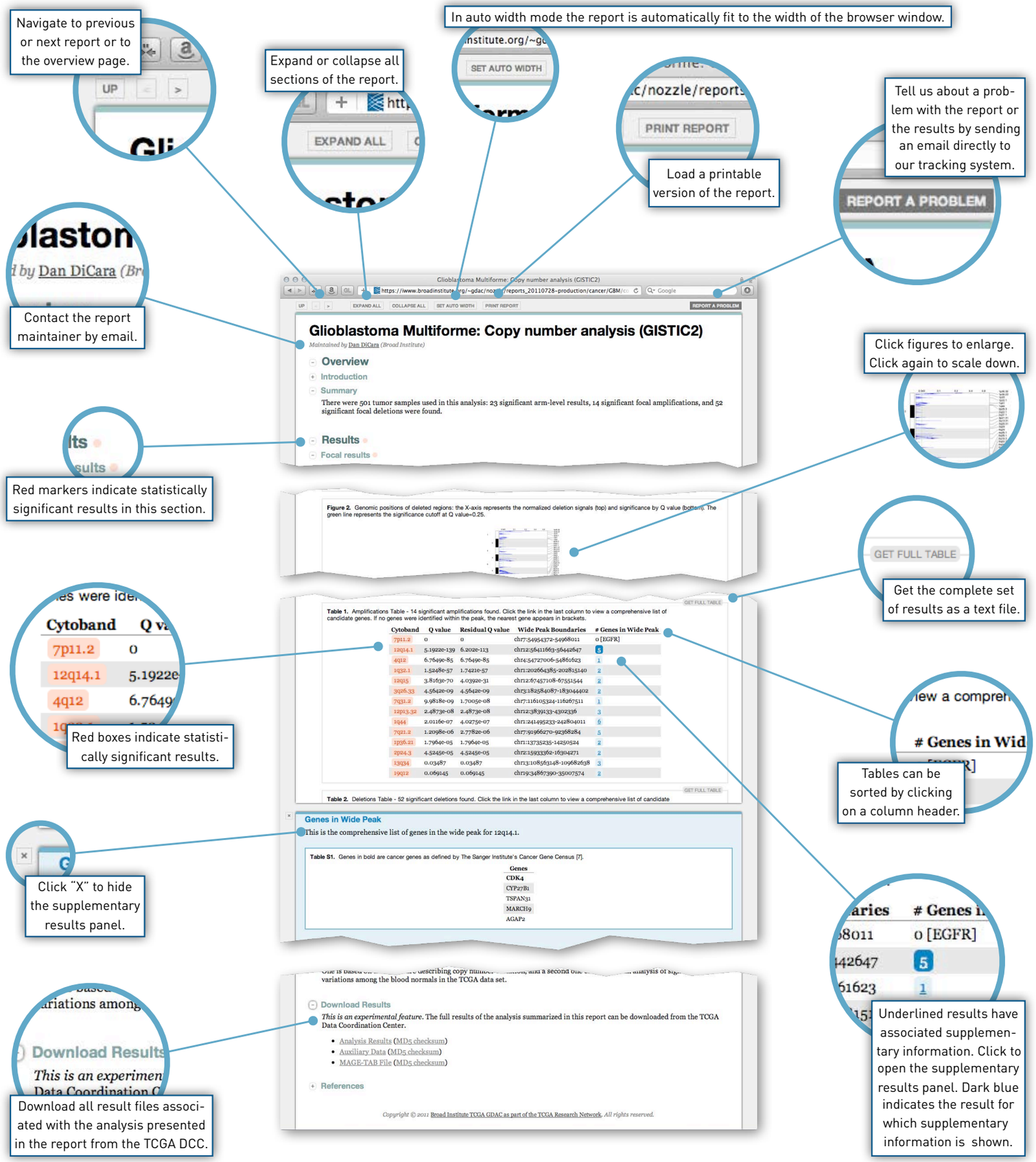
Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | Significantly mutated genes ($q \leq 0.1$): 24
- *Clustering Analyses*
  - **Clustering of mRNA expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of *565* samples using the *1500* most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of mRNA expression: consensus hierarchical**
    View Report | The *1500* most variable genes were selected. Consensus average linkage hierarchical clustering of *565* samples and *1500* genes identified *3* subtypes with the stability of the clustering increasing for $k = 2$ to $k = 8$ and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of Methylation: consensus NMF**
    View Report | The *1229* most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distriution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of *551* samples and *1229* genes identified *6* subtypes with the stability of the clustering increasing for $k = 2$ to $k = 8$ and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to *150* most variable miRs. Consensus NMF clustering of *564* samples and *150* miRs identified *3* subtypes with the stability of the clustering increasing for $k = 2$ to $k = 8$ and the average silhouette width calculation for selecting the robust clusters.

Firehose Reports | At-a-Glance

BROAD INSTITUTE

➜ Reports are compatible with Firefox 4+, Chrome 12+, Safari 5+, Opera 11+ and Internet Explorer 9+.



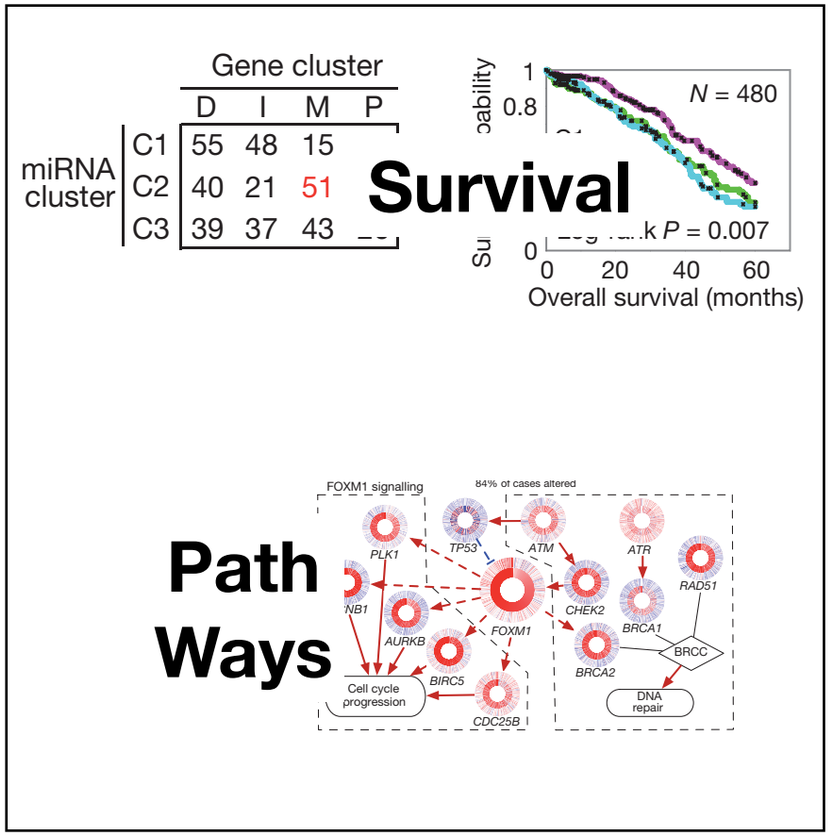
Organized like a paper

- Overview ("Abstract")
- Results
- Methods & Data

With Browser Convenience

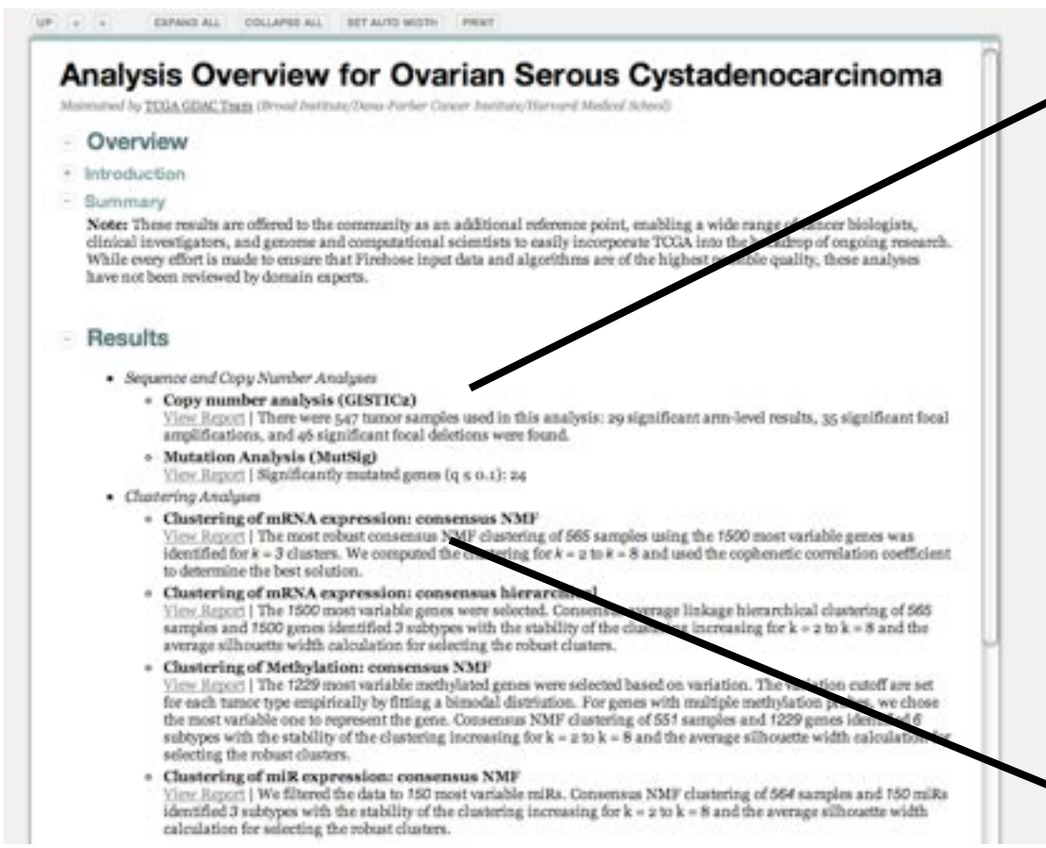
- Dynamic zooming
- And navigation
- View partial or full data
- Easily printable
- Built-in bug reporting
- No HTML coding: just R

- Aim is to enable readers (PIs, bench bios, clinical trialists)
- To quickly take pulse of TCGA for given tumor type(s)
- With just a few glances at common representational figures
- Not deep head-scratching



Low hanging fruit

Including Survival & Pathway analyses



UP + + EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by TCGA GDAC Team (Broad Institute/Dana-Farber Cancer Institute/Harvard Medical School)

Overview

Introduction

Summary

Note: These results are offered to the community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the background of ongoing research. While every effort is made to ensure that Firehose input data and algorithms are of the highest possible quality, these analyses have not been reviewed by domain experts.

Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC2)**
    View Report | There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.
  - **Mutation Analysis (MutSig)**
    View Report | Significantly mutated genes (q ≤ 0.1): 24
- *Clustering Analyses*
  - **Clustering of mRNA expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.
  - **Clustering of mRNA expression: consensus hierarchical**
    View Report | The 1500 most variable genes were selected. Consensus average linkage hierarchical clustering of 565 samples and 1500 genes identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of Methylation: consensus NMF**
    View Report | The 1229 most variable methylated genes were selected based on variation. The variation cutoff are set for each tumor type empirically by fitting a bimodal distribution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of 551 samples and 1229 genes identified 4 6 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.
  - **Clustering of miR expression: consensus NMF**
    View Report | We filtered the data to 150 most variable miRs. Consensus NMF clustering of 564 samples and 150 miRs identified 3 subtypes with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.

Ovarian Serous Cystadenocarcinoma: Copy number analysis (GISTIC2)

Maintained by Dan DiCara (Broad Institute)

Overview

Introduction

Summary

There were 547 tumor samples used in this analysis: 29 significant arm-level results, 35 significant focal amplifications, and 46 significant focal deletions were found.

Results

Focal results

Figure 1. Genomic positions of amplified regions: the X-axis represents the normalized amplification signals (top) and significance by Q value (bottom). The green line represents the significance cutoff at Q value=0.25.

GET FULL TABLE

Table 1. Amplifications Table - 35 significant amplifications found. Click the link in the last column to view a comprehensive list of candidate genes. If no genes were identified within the peak, the nearest gene appears in brackets.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
8q24.21	2.645e-77	2.645e-77	chr8:128574848-129810279	5
19q12	1.8147e-87	8.4949e-76	chr19:34947990-35023082	1
3q26.2	1.07220e-60	1.07220e-60	chr3:170905217-170925258	0 [MECOM]

Ovarian Serous Cystadenocarcinoma: Clustering of mRNA expression: consensus NMF

Maintained by Robert Zupko (Broad Institute)

Overview

Introduction

Summary

The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for $k = 3$ clusters. We computed the clustering for $k = 2$ to $k = 8$ and used the cophenetic correlation coefficient to determine the best solution.

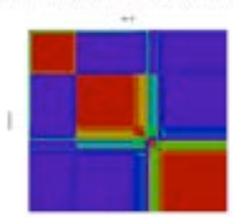
Results

Gene expression patterns of molecular subtypes

Consensus and correlation matrix

GET HIGH-RES IMAGE

Figure 2. The consensus matrix after clustering shows 3 clusters with limited overlap between clusters.



Diverse set of front-line analyses
Point/Click From Desktop
No passwords

Ovarian Serous Cystadenocarcinoma: Copy number analysis (GISTIC2)

Maintained by Dan DiCara *(Broad Institute)*

Overview

Introduction

Summary

There were 558 tumor samples used in this analysis: 29 significant arm-level results, 34 significant focal amplifications, and 47 significant focal deletions were found.

Results

Focal results

Arm-level results

Methods & Data

Input

GISTIC

Download Results

This is an experimental feature. The full results of the analysis summarized in this report can be downloaded from the TCGA Data Coordination Center.

- Analysis Results (MD5 checksum)
- Auxiliary Data (MD5 checksum)
- MAGE-TAB File (MD5 checksum)

References

[1] Beroukhim et al, Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma, *Proc Natl Acad Sci U S A.* **Vol. 104**:50 (2007)

[2] GISTIC version 1

[3] Mermel et al, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, *Genome Biology* **Vol. 12**:4 (2011)

[4] GISTIC version 2

[5] Beroukhim et al., The landscape of somatic copy-number alteration across human cancers, *Nature* **Vol. 463**:7283 (2010)

[6] McCarroll, S. A. et al., Integrated detection and population-genetic analysis of SNPs and copy number variation, *Nat Genet* **Vol. 40**(10):1166-1174 (2008)

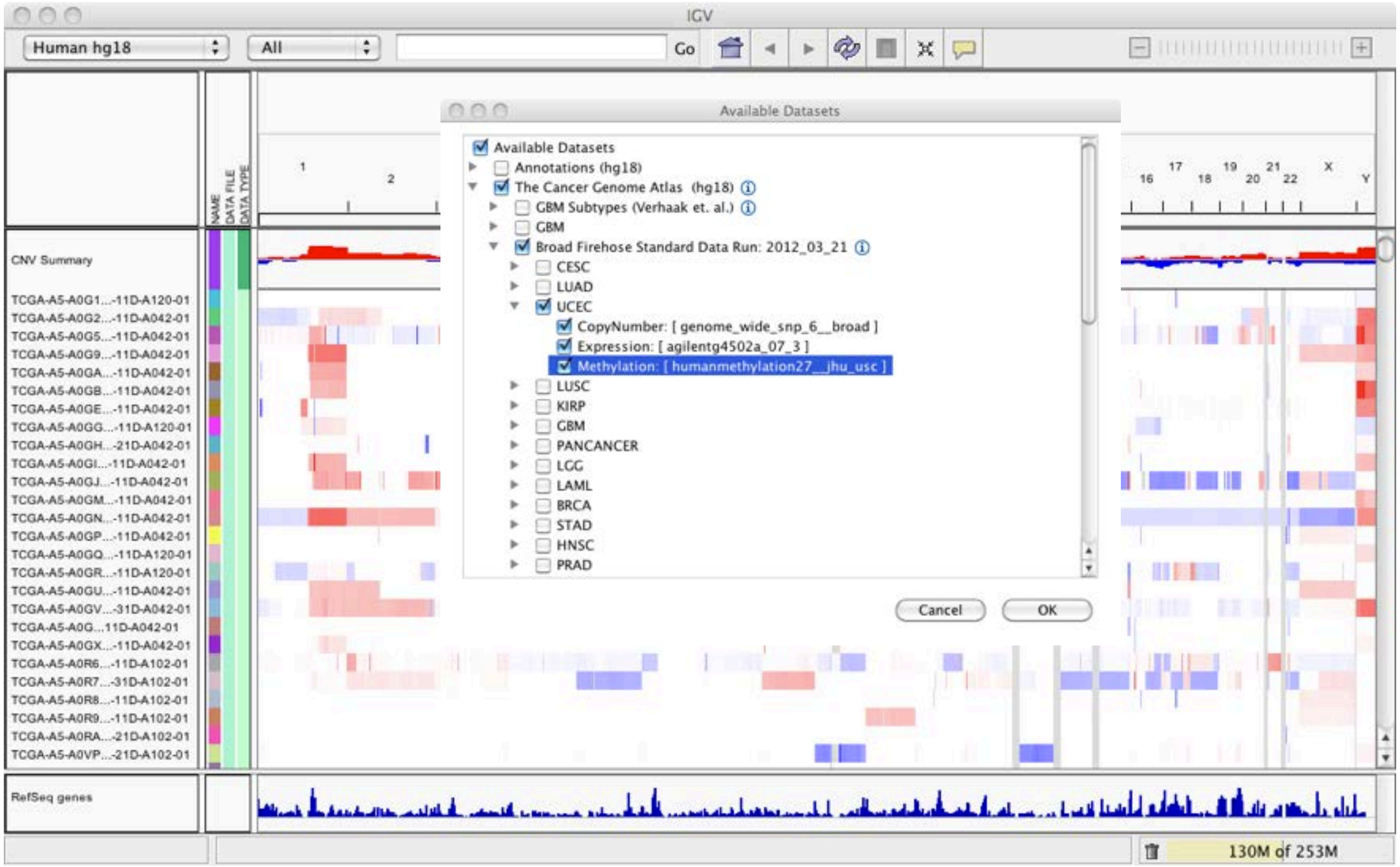
[7] The Sanger Institute: Cancer Gene Census

Directly Linked to Data Used

And References

Automatically Generated for All Runs

PRE-LOADED IN IGV, TOO



INTEGRATIVE GENOMICS VIEWER : **www.broadinstitute.org/igv/**

Interactive or Programmatic Download

firehose_get *v0.3.2*

```
Usage: firehose_get [flags]  RunType  Date  [tumor_type, ... ]

Two arguments are required; the first must be one of

   analyses    |   stddata

while the second must EITHER be a date (in YYYY_MM_DD form) of an
existing GDAC run of the given type OR 'latest'.  An optional third,
fourth etc argument may be specified to prune the retrieval, given
as a subset of these case-insensitive TCGA tumor type abbreviations:

  BLCA BRCA CESC COADREAD DLBC GBM HNSC KIRC KIRP LAML LGG LIHC
  LNNH LUAD LUSC OV PAAD PRAD SKCM STAD THCA UCEC PANCANCER

Note that as a convenience 'analysis' and 'data' are accepted as
synonyms for the 'analyses' and 'stddata' run types

Flags:

  -b | -batch          do not prompt: assume YES answer to all queries
  -e | -echo           show commands that would be run, but do nothing
  -h | -help | --help this message
  -l | -log            write output to log file, instead of stdout
  -r | -runs           display list of all available Firehose runs
  -t | -tasks <list>   further prune the set of archives retrieved, by
                       downloading ONLY the tasks (pipelines) whose
                       names match the given space-delimited list of
                       patterns; matching is performed with glob-style
                       wildcards; when no pattern list is given firehose_get
                       will display all tasks in the selected run
                       NOTE: not all tasks will execute for all tumor
                             sets; what tasks are run depends upon the
                             data available for that tumor type
  -v                   display the version of firehose_get
  -x                   debugging: turn on bash set -x (warning: very verbose)
```

% firehose_get -runs

Run	At_DCC	Available_From_Broad_GDAC
stddata__2011_10_26	yes	no
stddata__2011_11_15	yes	no
stddata__2011_11_28	yes	no
stddata__2011_12_06	yes	no
stddata__2011_12_30	yes	no
stddata__2012_01_10	yes	no
stddata__2012_01_24	yes	no
stddata__2012_02_17	yes	yes
stddata__2012_03_06	yes	yes
stddata__2012_03_21	yes	yes
stddata__2012_04_12	yes	yes
stddata__2012_04_25	yes	yes
stddata__2012_05_15	yes	yes
stddata__2012_05_25	no	yes
analyses__2010_12_23	yes	no
analyses__2011_01_14	yes	no
analyses__2011_02_17	yes	no
analyses__2011_03_27	yes	no
analyses__2011_04_21	yes	no
analyses__2011_05_25	yes	no
analyses__2011_07_28	yes	no
analyses__2011_09_21	yes	no
analyses__2011_10_26	yes	no
analyses__2011_11_28	yes	no
analyses__2011_12_30	yes	no
analyses__2012_01_24	yes	no
analyses__2012_02_17	yes	yes
analyses__2012_03_21	yes	yes
analyses__2012_04_25	yes	yes
analyses__2012_05_25	no	yes

Quickly discern what versioned runs have been performed.

```
% firehose_get -tasks analyses 2012_05_25
```

```
CopyNumber_GeneBySample
CopyNumber_Gistic2
Correlate_CopyNumber_vs_miR
Correlate_CopyNumber_vs_mRNA
Correlate_CopyNumber_vs_mRNAseq
Correlate_Methylation_vs_mRNA
Methylation_Clustering_CNMF
miRseq_Clustering_CNMF
miRseq_Clustering_Consensus
miRseq_Preprocess
miR_Clustering_CNMF
miR_Clustering_Consensus
miR_FindDirectTargets
mRNAseq_Clustering_CNMF
mRNAseq_Clustering_Consensus
mRNAseq_Preprocess
mRNA_Clustering_CNMF
mRNA_Clustering_Consensus
mRNA_Preprocess_Median
Mutation_Assessor
Mutation_Significance
Pathway_FindEnrichedGenes
Pathway_Paradigm_Expression
Pathway_Paradigm_Expression_CopyNumber
RPPA_Clustering_CNMF
RPPA_Clustering_Consensus
```

Or what those runs contain.

```
% firehose_get -tasks Methylation analyses 2012_05_25 OV GBM UCEC
```

```
You've asked to download archives for the following tasks

     Methylation

run against the tumor datasets

      OV GBM UCEC

from the analyses__2012_05_25 Firehose run. If this is correct,
shall we continue with download? (y|yes|n|no) [no] y
```

Pick pieces of a run.

```
% firehose_get  stddata  latest
```

```
You've asked to download archives for the following tumor datasets

     BLCA BRCA CESC COADREAD DLBC GBM HNSC KIRC KIRP LAML LGG LIHC LNNH LUAD LUSC OV PAAD PRAD SKCM STAD THCA UCEC PANCANCER

from the stddata__2012_05_25 Firehose run. If this is correct,
shall we continue with download? (y|yes|n|no) [no] y

Attempting to retrieve data for Broad GDAC run stddata__2012_05_25 ...
--2012-06-28 08:42:23--  http://gdac.broadinstitute.org/runs/stddata__2012_05_25/data/BLCA/
     0K                                                      100% 17.6M=0s
     0K                                                      100%  309M=0s
--2012-06-28 08:42:23--  http://gdac.broadinstitute.org/runs/stddata__2012_05_25/data/BLCA/20120525/
     0K                                                       251M=0s
```

Or everything latest ... et cetera.

FIREHOSE
Broad GDAC

Offered to community as an additional reference point, enabling a wide range of cancer biologists, clinical investigators, and genome and computational scientists to easily incorporate TCGA into the backdrop of ongoing research.

FIREHOSE
Broad GDAC

- Signficant progress to date
- > 1000 data & analyses pipelines run per month
- Packaged for easy browse/download
- One-stop accessibility
- Historical comparison: all runs, not only latest
- Persistent at DCC
- Citable by exact aggregate version
- TCGA manuscript provenance/freeze
- But plenty of challenges remain, including

  - ✓ Time lag (race condition) for new data
  - ✓ Time lag for latest analyses (e.g. meth27 & 450)
  - ✓ Incorporating batch effects upfront
  - ✓ QC kilopipeline per month

Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad
Michael Noble
Douglas Voet
Gordon Saksena
Dan DiCara
Kristian Cibulskis
Juok Cho
Rui Jing
Michael Lawrence
Lee Lichtenstein
Pei Lin
Spring Liu
William Mallard
Aaron McKenna
Sachet Shukla
Raktim Sinha
Andrey Sivachenko
Carrie Sougnez
Petar Stojanov
Lihua Zou
Hailei Zhang
Robert Zupko

Belfler-DFCI/MDACC
Yonghong Xiao
Juinhua Zhang
Terrence Wu

IGV & GenePattern teams @ Broad
Jill Mesirov
Michael Reich
Peter Carr
Marc-Danie Nazaire
Jim Robinson
Helga Thorvaldsdottir

Harvard
Peter Park
Nils Gehlenborg
Semin Lee
Richard Park

Matthew Meyerson
Todd Golub
Eric Lander

THE CANCER GENOME ATLAS

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center Making Cancer History®