



# CPTAC Introduction To Firehose

Michael S. Noble  
The Broad Institute of MIT & Harvard

CPTAC Steering Committee Meeting  
Houston, Texas

April 25, 2012



# Goals of This Short Talk

---

I. Purpose of Firehose in TCGA

II. Outline major operational products: data & analysis runs

III. How & Where to retrieve them

IV. Provide sense of potential value to CPTAC

# I. Purpose

---



Born of the desire to systematize analyses from The Cancer Genome Atlas (TCGA) 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 ...

---

# Because The Bad Old Days ...

---

Of solitary, manual experimentation ...

```
% create a folder
```

```
% download data.from.some.where
```

```
% run_your_computational_analysis
```

# 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, ...

# 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

2012\_03\_06 stddata Run

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	MAF
BLCA	89	65	35	59	0	0	0	54	28
BRCA	864	844	781	808	529	751	0	781	507
CESC	99	12	36	0	0	0	0	8	0
COADREAD	591	590	565	585	224	83	0	255	224
DLBC	10	0	0	0	0	0	0	0	0
GBM	596	561	537	287	542	0	491	0	276
HNSC	294	227	165	292	0	13	0	89	0
KIRC	502	502	489	500	72	469	0	463	327
KIRP	129	84	43	36	16	14	0	16	0
LAML	202	200	0	192	0	179	0	187	199
LGG	144	117	80	0	27	0	0	30	0
LIHC	84	47	53	0	0	17	0	28	0
LNNH	2	0	0	0	0	0	0	0	0
LUAD	371	272	205	325	32	0	0	95	147
LUSC	290	269	211	282	154	220	0	202	178
OV	592	580	547	551	568	0	564	46	316
PAAD	38	0	14	0	0	0	0	0	0
PRAD	153	0	82	0	0	0	0	63	0
SKCM	253	0	0	0	0	0	0	0	0
STAD	168	163	134	118	0	58	0	125	0
THCA	274	73	85	0	0	0	0	45	0
UCEC	462	392	363	373	54	266	0	359	239
PANCANCER	6207	4998	4425	4408	2218	2070	1055	2846	2441

March 2012 Samples In Firehose



# Doesn't Scale to TCGA

2012\_04\_12 stddata Run

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

# Doesn't Scale to TCGA

2012\_04\_12 stddata Run

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

+821  
CopyNumber

April 2012

# Doesn't Scale to TCGA

2012\_04\_12 stddata Run

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

+821  
CopyNumber

April 2012

+917  
Methylation

# Doesn't Scale to TCGA

New datatype column  
+2087 protein samples

2012\_04\_12 stddata Run

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

+821  
CopyNumber

April 2012

+917  
Methylation

## II. So Firehose Automatically Generates

---

- 1 Regular package of standard analyses results (~monthly)\*  
*For vetted algorithms: GISTIC, MutSig, CNMF, ...*
- 2 From version-stamped, standardized datasets  
*Generated at Broad, precursor to automated analyses*

## II. So Firehose Automatically Generates

---

1 Regular package of standard analyses results (~monthly)\*

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

2 From version-stamped, standardized datasets

*Generated at Broad, precursor to automated analyses*

\* Companioned with biologist-friendly reports

Analyses Pipelines: 26 x 23 tumor sets / month = 598

Standardized Datasets: 273 platforms (in 23 tumorsets) x 2/month = 546

# But why Firehose ...

---

Home Query the Data Download Data Tools About the Data

Home

## TCGA Data Portal Overview

We provide 3 ways to download data: The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high-throughput sequencing analysis of the tumor genomes.

[Query the Data ▶](#) [Download Data ▶](#)

Search summarized data for genes, patients and pathways Choose from three ways to download data

Available Cancer Types	# Patients with Samples	# Downloadable Tumor Samples	Date Last Updated (mm/dd/yy)
<a href="#">Acute Myeloid Leukemia [LAML]</a>	202	200	02/22/12
<a href="#">Bladder Urothelial Carcinoma [BLCA]</a>	89	78	03/20/12

... when TCGA data portal already exists?

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

---

No aggregate versioning

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?



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

---

No aggregate versioning

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

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

---

No aggregate versioning

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 generate analyses at all

What if I just want to view OV Gistic (CN) peaks?

Or peek at an expression or methylation cluster?

You might otherwise need to ...

---

Spend weeks obtaining protected data credentials

Or becoming a TCGA data guru

And still more time, mastering the analytics

You might otherwise need to ...

---

Spend weeks obtaining protected data credentials

Or becoming a TCGA data guru

And still more time, mastering the analytics

AGAIN: complexity & volume preclude  
this approach for many individuals

III. Ok, where do I start?

---

<http://gdac.broadinstitute.org>

# III. Ok, where do I start?

---

<http://gdac.broadinstitute.org>

2012\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download
<a href="#">BLCA</a>	8	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">BRCA</a>	16	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">CESC</a>	7	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">COADREAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">GBM</a>	21	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">HNSC</a>	12	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRC</a>	16	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRP</a>	12	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LAML</a>	7	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LGG</a>	8	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LHC</a>	8	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUAD</a>	16	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUSC</a>	25	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">OV</a>	23	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PAAD</a>	3	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PRAD</a>	5	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">SKCM</a>	1	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">STAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">THCA</a>	7	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">UCEC</a>	16	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PANCANCER</a>	34	<a href="#">85%</a>	<a href="#">Open</a> <a href="#">Protected</a>

Data  
Dashboard

2012\_03\_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
<a href="#">BRCA</a>	23	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">COADREAD</a>	23	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">GBM</a>	21	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LGG</a>	14	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUSC</a>	23	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">OV</a>	24	<a href="#">100%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRC</a>	22	<a href="#">96%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUAD</a>	22	<a href="#">96%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">UCEC</a>	22	<a href="#">96%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">STAD</a>	14	<a href="#">93%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRP</a>	16	<a href="#">89%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PRAD</a>	7	<a href="#">88%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">THCA</a>	7	<a href="#">88%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LAML</a>	11	<a href="#">85%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">BLCA</a>	7	<a href="#">78%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">HNSC</a>	7	<a href="#">78%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LHC</a>	7	<a href="#">78%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">CESC</a>	6	<a href="#">60%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PAAD</a>	3	<a href="#">60%</a>	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PANCANCER</a>	11	<a href="#">58%</a>	<a href="#">Open</a> <a href="#">Protected</a>

Analysis  
Dashboard



# Standardized Data Dashboard

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**

## 2012\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download	
<a href="#">BLCA</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">BRCA</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">CESC</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">COADREAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">GBM</a>	21	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">HNSC</a>	12	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRC</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRP</a>	12	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LAML</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LGG</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LHC</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUAD</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUSC</a>	25	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">OV</a>	23	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PAAD</a>	3	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PRAD</a>	5	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">SKCM</a>	1	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">STAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">THCA</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">UCEC</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PANCANCER</a>	34	<a href="#">85%</a>	<a href="#">Open</a>	<a href="#">Protected</a>

# Standardized Data Dashboard

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**

2012\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download	
<a href="#">BLCA</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">BRCA</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">CESC</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">COADREAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">GBM</a>	21	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">HNSC</a>	12	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRC</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRP</a>	12	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LAML</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LGG</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LHC</a>	8	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUAD</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUSC</a>	25	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">OV</a>	23	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PAAD</a>	3	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PRAD</a>	5	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">SKCM</a>	1	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">STAD</a>	14	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">THCA</a>	7	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">UCEC</a>	16	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PANCANCER</a>	34	<a href="#">85%</a>	<a href="#">Open</a>	<a href="#">Protected</a>

Versioned

Helps BABEL problem

promoting agreement  
across centers on  
sample counts

Great starting point for  
aggregated TCGA data

→ ICGC, too!



Fine, but where are those easy analytics?

---

GISTIC copynumber amp/del peaks?

Or methylation/expression clusters?

Or mutation significance tables?

## 2012\_03\_21 analyses Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

AnalysisReport	# Pipelines	% Successful	Download	
<a href="#">BRCA</a>	23	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">COADREAD</a>	23	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">GBM</a>	21	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LGG</a>	14	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUSC</a>	23	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">OV</a>	24	<a href="#">100%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRC</a>	22	<a href="#">96%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LUAD</a>	22	<a href="#">96%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">UCEC</a>	22	<a href="#">96%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">STAD</a>	14	<a href="#">93%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">KIRP</a>	16	<a href="#">89%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PRAD</a>	7	<a href="#">88%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">THCA</a>	7	<a href="#">88%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LAML</a>	11	<a href="#">85%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">BLCA</a>	7	<a href="#">78%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">HNSC</a>	7	<a href="#">78%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">LIHC</a>	7	<a href="#">78%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">CESC</a>	6	<a href="#">60%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PAAD</a>	3	<a href="#">60%</a>	<a href="#">Open</a>	<a href="#">Protected</a>
<a href="#">PANCANCER</a>	11	<a href="#">58%</a>	<a href="#">Open</a>	<a href="#">Protected</a>

# Analyses Dashboard

## 2012\_03\_21 analyses Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNAseq	miR	miRseq	MAF
BLCA	89	58	35	38	0	0	0	54	0
BRCA	864	844	781	808	529	751	0	781	507
CESC	99	12	36	0	0	0	0	8	0
COADREAD	591	586	565	236	224	83	0	255	224
DLBC	10	0	0	0	0	0	0	0	0
GBM	596	561	537	287	542	0	491	0	276
HNSC	294	206	165	0	0	13	0	89	0
KIRC	502	502	489	391	72	469	0	463	327
KIRP	129	84	43	36	16	14	0	16	0
LAML	202	200	0	192	0	179	0	187	199
LGG	119	103	80	0	27	0	0	30	0
LHC	84	47	53	0	0	17	0	28	0
LNNH	2	0	0	0	0	0	0	0	0
LUAD	353	269	205	127	32	0	0	95	147
LUSC	283	266	211	133	154	220	0	202	178
OV	592	580	547	551	568	0	564	46	316
PAAD	38	0	14	0	0	0	0	0	0
PRAD	153	0	82	0	0	0	0	63	0
SKCM	240	0	0	0	0	0	0	0	0
STAD	149	148	134	118	0	58	0	125	0
THCA	251	54	85	0	0	0	0	45	0
UCEC	462	392	363	373	54	266	0	359	239
Totals	6102	4912	4425	3290	2218	2070	1055	2846	2413

AnalysisReport	# Pipelines	% Successful	Download
<a href="#">BRCA</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">COADREAD</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">GBM</a>	21	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LGG</a>	14	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUSC</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">OV</a>	24	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRC</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUAD</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">UCEC</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">STAD</a>	14	93%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRP</a>	16	89%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PRAD</a>	7	88%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">THCA</a>	7	88%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LAML</a>	11	85%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">BLCA</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">HNSC</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LIHC</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">CESC</a>	6	60%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PAAD</a>	3	60%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PANCANCER</a>	11	58%	<a href="#">Open</a> <a href="#">Protected</a>

Pipeline	Not Ready	Failed	Succeed
1 Aggregate_Clusters	0	0	1
2 Clinical_Aggregate_Tier1	0	0	1
3 Clinical_Pick_Tier1	0	0	1
4 CopyNumber_GeneBySample	0	0	1
5 CopyNumber_Gistic2	0	0	1
6 CopyNumber_Preprocess	0	0	1
7 Correlate_Clinical_vs_miR	0	0	1
8 Correlate_Clinical_vs_Molecular_Signatures	0	0	1
9 Correlate_Clinical_vs_mRNA	0	0	1
10 Correlate_Clinical_vs_Mutation	0	0	1
11 Correlate_CopyNumber_vs_miR	0	0	1
12 Correlate_CopyNumber_vs_mRNA	0	0	1
13 Correlate_GenomicEvents	0	0	1
14 Correlate_Methylation_vs_mRNA	0	0	1
15 miR_Clustering_CNMF	0	0	1
16 miR_Clustering_Consensus	0	0	1
17 miR_FindDirectTargets	0	0	1
18 mRNA_Clustering_CNMF	0	0	1
19 mRNA_Clustering_Consensus	0	0	1
20 mRNA_Preprocess_Median	0	0	1
21 Mutation_Assessor	0	0	1
22 Mutation_Significance	0	0	1
23 Pathway_FindEnrichedGenes	0	0	1
24 Pathway_Paradigm	0	0	1
Total	0	0	24

20+ analyses per  
20+ tumorsets

# Dashboards Updated

## 2012\_02\_17 analyses Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#)

AnalysisReport	# Pipelines	% Successful	Download
<a href="#">OV</a>	24	100%	<a href="#">Archives</a>
<a href="#">BRCA</a>	23	100%	<a href="#">Archives</a>
<a href="#">COADREAD</a>	23	100%	<a href="#">Archives</a>
<a href="#">LUSC</a>	23	100%	<a href="#">Archives</a>
<a href="#">GBM</a>	21	100%	<a href="#">Archives</a>
<a href="#">LGG</a>	14	100%	<a href="#">Archives</a>
<a href="#">KIRC</a>	22	96%	<a href="#">Archives</a>
<a href="#">UCEC</a>	22	96%	<a href="#">Archives</a>
<a href="#">LUAD</a>	19	95%	<a href="#">Archives</a>
<a href="#">KIRP</a>	16	89%	<a href="#">Archives</a>
<a href="#">BLCA</a>	7	88%	<a href="#">Archives</a>
<a href="#">PRAD</a>	7	88%	<a href="#">Archives</a>
<a href="#">THCA</a>	7	88%	<a href="#">Archives</a>
<a href="#">LAML</a>	11	85%	<a href="#">Archives</a>
<a href="#">STAD</a>	11	85%	<a href="#">Archives</a>
<a href="#">HNSC</a>	7	78%	<a href="#">Archives</a>
<a href="#">LIHC</a>	7	78%	<a href="#">Archives</a>
<a href="#">PAAD</a>	3	60%	<a href="#">Archives</a>
<a href="#">CESC</a>	4	50%	<a href="#">Archives</a>
<a href="#">DLBC</a>	0	0%	
<a href="#">LNNH</a>	0	0%	
<a href="#">SKCM</a>	0	0%	

# Dashboards Updated

## 2012\_02\_17 analyses Run

Tables of Ingested Data: [HTML](#) [RML](#) [TSV](#)

AnalysisReport	# Pipelines	% Successful	Download
<a href="#">OV</a>	24	100%	<a href="#">Archives</a>
<a href="#">BRCA</a>	23	100%	<a href="#">Archives</a>
<a href="#">COADREAD</a>	23	100%	<a href="#">Archives</a>
<a href="#">LUSC</a>	23	100%	<a href="#">Archives</a>
<a href="#">GBM</a>	21	100%	<a href="#">Archives</a>
<a href="#">LGG</a>	14	100%	<a href="#">Archives</a>
<a href="#">KIRC</a>	22	96%	<a href="#">Archives</a>
<a href="#">UCEC</a>	22	96%	<a href="#">Archives</a>
<a href="#">LUAD</a>	19	95%	<a href="#">Archives</a>
<a href="#">KIRP</a>	16	89%	<a href="#">Archives</a>
<a href="#">BLCA</a>	7	88%	<a href="#">Archives</a>
<a href="#">PRAD</a>	7	88%	<a href="#">Archives</a>
<a href="#">THCA</a>	7	88%	<a href="#">Archives</a>
<a href="#">LAML</a>	11	85%	<a href="#">Archives</a>
<a href="#">STAD</a>	11	85%	<a href="#">Archives</a>
<a href="#">HNSC</a>	7	78%	<a href="#">Archives</a>
<a href="#">LIHC</a>	7	78%	<a href="#">Archives</a>
<a href="#">PAAD</a>	3	60%	<a href="#">Archives</a>
<a href="#">CESC</a>	4	50%	<a href="#">Archives</a>
<a href="#">DLBC</a>	0	0%	
<a href="#">LNNH</a>	0	0%	
<a href="#">SKCM</a>	0	0%	

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 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 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

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

**Navigation:** Navigate to previous or next report or to the overview page.

**Layout:** In auto width mode the report is automatically fit to the width of the browser window.

**Interactions:** Expand or collapse all sections of the report. Load a printable version of the report. Tell us about a problem with the report or the results by sending an email directly to our tracking system.

**Reporting:** Contact the report maintainer by email.

**Significance:** Red markers indicate statistically significant results in this section. Red boxes indicate statistically significant results.

**Figure 2:** Genomic positions of deleted regions. The X-axis represents the normalized deletion signals (top) and significance by Q value (bottom). The green line represents the significance cutoff at Q value=0.25.

**Table 1:** Amplifications Table - 14 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
<u>7p11.2</u>	0	0	chr7:54954372-54968011	0 [EGFR]
<u>12q14.1</u>	5.19228	6.2028-113	chr12:36411663-36442647	<b>5</b>
<u>4q12</u>	6.76498	6.76498-85	chr4:54727006-54861623	<b>1</b>
<u>13q32.1</u>	1.32488	1.74218-57	chr13:202664385-202815140	<b>2</b>
<u>12q15</u>	3.81638	4.03928-31	chr12:67457108-67551544	<b>2</b>
<u>3p26.33</u>	4.56428	4.56428-09	chr3:182584087-183044402	<b>2</b>
<u>7q31.2</u>	9.98188	1.70058-08	chr7:116105324-116267511	<b>1</b>
<u>12p13.32</u>	2.48738	2.48738-08	chr12:38391333-4302336	<b>3</b>
<u>13q44</u>	2.01188	4.02758-07	chr13:241495233-242804011	<b>6</b>
<u>7q21.2</u>	1.20988	2.77828-06	chr7:39366270-422804011	<b>5</b>
<u>13q32.1</u>	1.79648	1.79648-05	chr13:13735235-14250524	<b>2</b>
<u>2p24.3</u>	4.32488	4.52488-05	chr2:15933362-16304271	<b>2</b>
<u>13q34</u>	0.03487	0.03487	chr13:108563148-109682638	<b>3</b>
<u>19q12</u>	0.069145	0.069145	chr19:34867390-35007574	<b>2</b>

**Table 2:** Deletions Table - 52 significant deletions found. Click the link in the last column to view a comprehensive list of candidate genes.

**Genes in Wide Peak:** This is the comprehensive list of genes in the wide peak for 12q14.1.

Genes
<b>CDK4</b>
<b>CTP27B1</b>
<b>TSPAN31</b>
<b>MARCKS19</b>
<b>AGAP2</b>

**Table S1:** Genes in bold are cancer genes as defined by The Sanger Institute's Cancer Gene Census [7].

**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.

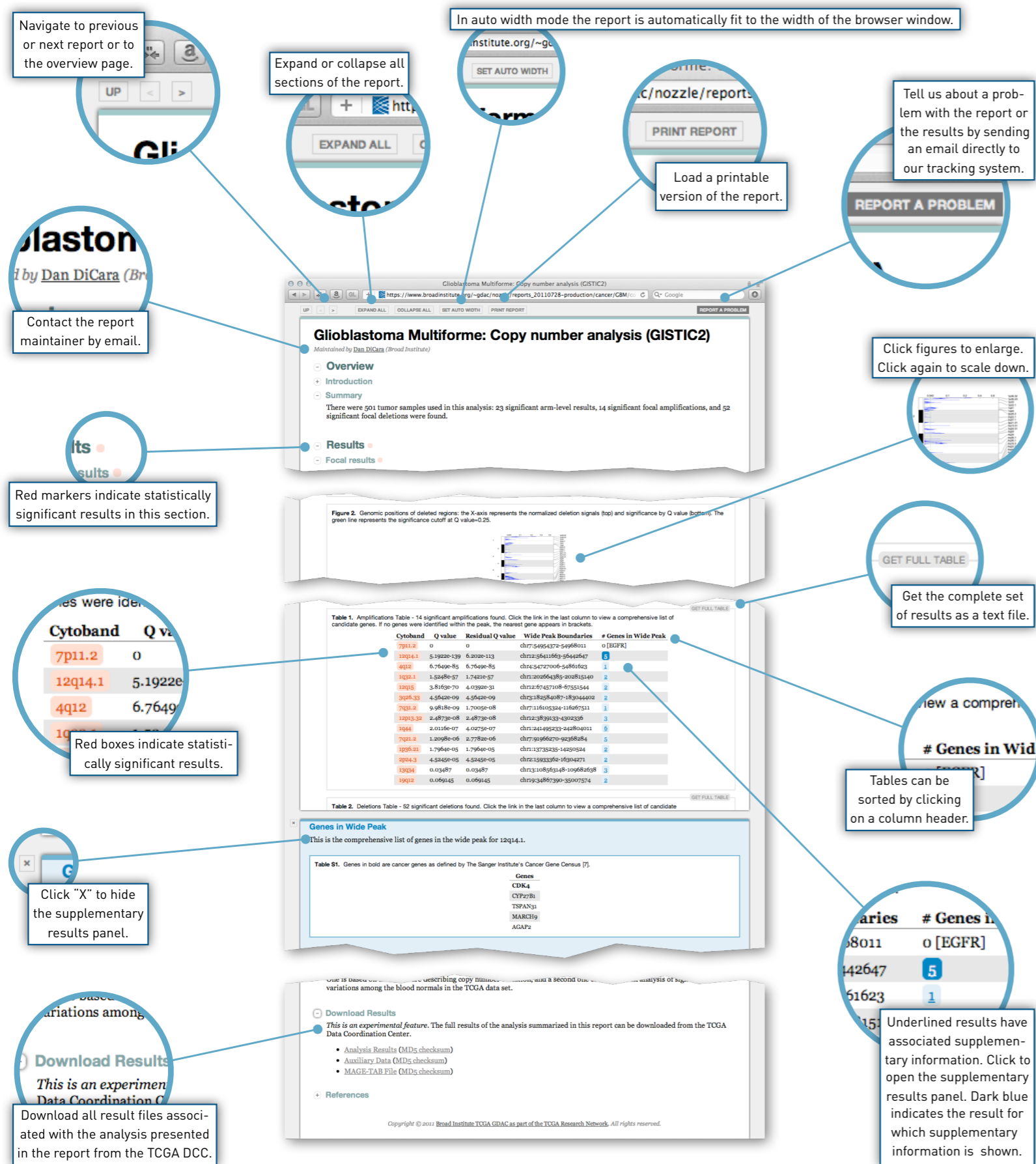
- Analysis Results (MD5 checksum)
- Auxiliary Data (MD5 checksum)
- MAGE-TAB File (MD5 checksum)

**References:**

**Tables:** Tables can be sorted by clicking on a column header.

**Supplementary Information:** Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

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



**Navigation:**

- Navigate to previous or next report or to the overview page.
- Expand or collapse all sections of the report.
- In auto width mode the report is automatically fit to the width of the browser window.
- Load a printable version of the report.
- Tell us about a problem with the report or the results by sending an email directly to our tracking system.
- Contact the report maintainer by email.

**Results and Data:**

- Red markers indicate statistically significant results in this section.
- Red boxes indicate statistically significant results.
- Click figures to enlarge. Click again to scale down.
- Get the complete set of results as a text file.
- Tables can be sorted by clicking on a column header.
- Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

**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)

## Organized like a paper

- Overview (“Abstract”)
- Results
- Methods & Data

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

**Navigation and Convenience:**

- Navigate to previous or next report or to the overview page.
- Expand or collapse all sections of the report.
- In auto width mode the report is automatically fit to the width of the browser window.
- Load a printable version of the report.
- Tell us about a problem with the report or the results by sending an email directly to our tracking system.
- Contact the report maintainer by email.
- Click "X" to hide the supplementary results panel.
- Download all result files associated with the analysis presented in the report from the TCGA DCC.

**Data and Results:**

- Red markers indicate statistically significant results in this section.
- Red boxes indicate statistically significant results.
- Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

**Table 1: Amplifications Table - 14 significant amplifications found.**

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
7p11.2	0	0	chr7:54954372-54968011	0 [EGFR]
12q14.1	5.1922e-09	6.202e-113	chr12:56411663-56442647	5
4q12	6.7649e-85	6.7649e-85	chr4:54727006-54861623	1
13q32.1	1.3248e-57	1.7421e-57	chr13:202664385-202815140	2
12q15	3.8163e-70	4.0392e-31	chr12:67457108-67551544	2
3p26.33	4.5642e-09	4.5642e-09	chr3:182584087-183044402	2
7q31.2	9.9818e-09	1.7005e-08	chr7:116103324-116267511	1
12p13.32	2.4873e-08	2.4873e-08	chr12:38391333-4302336	3
10q44	2.0116e-07	4.0275e-07	chr10:241495233-242804011	6
7q21.2	1.2098e-06	2.7782e-06	chr7:9266270-9268284	5
11p15.5	1.7964e-05	1.7964e-05	chr11:13735235-14250524	2
2p24.3	4.3245e-05	4.3245e-05	chr2:15933362-16304271	2
13q34	0.03487	0.03487	chr13:108563148-109682638	3
19q12	0.059145	0.059145	chr19:34867390-35007574	2

**Table 2: Deletions Table - 52 significant deletions found.**

**Table S1: Genes in bold are cancer genes as defined by The Sanger Institute's Cancer Gene Census [7].**

Genes
<b>CDK4</b>
<b>CTP27B1</b>
<b>TSPAN31</b>
<b>MARCKS19</b>
<b>AGAP2</b>

## 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



## Analysis Overview for Ovarian Serous Cystadenocarcinoma

Maintained by [TCGA GDHC 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 distribution. For genes with multiple methylation probes, we chose the most variable one to represent the gene. Consensus NMF clustering of 557 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.

## 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.



**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
<a href="#">8q24.21</a>	2.645e-77	2.645e-77	chr8:128574848-129810279	<a href="#">5</a>
<a href="#">19q12</a>	1.8147e-87	8.4949e-76	chr19:34947990-35023682	<a href="#">1</a>
<a href="#">3q26.2</a>	1.0722e-60	1.0722e-60	chr3:170905217-170923258	<a href="#">0</a> [MECOM]

## Ovarian Serous Cystadenocarcinoma: Clustering of mRNA expression: consensus NMF

Maintained by [Robert Zepko](#) (Broad Institute)

### Overview

#### Introduction

#### Summary

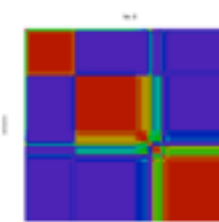
The most robust consensus NMF clustering of 565 samples using the 1500 most variable genes was identified for  $k = 3$  clusters. We computed the clustering for  $k = 2$  to  $k = 8$  and used the cophenetic correlation coefficient to determine the best solution.

### Results

#### Gene expression patterns of molecular subtypes

#### Consensus and correlation matrix

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



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 backbone 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 analysis via 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 variance 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 565 samples and 1229 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 miR expression: consensus NMF**  
[View Report](#) | We filtered the data to 150 most variable miRs. Consensus NMF clustering of 565 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.

Point/Click from browser  
Directly from Broad GDAC site  
No passwords  
Linked directly to downloadable data

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.

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.

- Aim is to quickly enable readers
- To take pulse of pipelines for given tumor type(s)
- By just glancing at common representational figures
- Not deep head-scratching
- Or weeks of data wrangling

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.

- Aim is to quickly enable readers
- To take pulse of pipelines for given tumor type(s)
- By just glancing at common representational figures
- Not deep head-scratching
- Or weeks of data wrangling

**VERY** low hanging fruit

# PRE-LOADED IN IGV, TOO

IGV

Human hg18 All Go

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

CNV Summary

TCGA-A5-A0G1...-11D-A120-01  
TCGA-A5-A0G2...-11D-A042-01  
TCGA-A5-A0G5...-11D-A042-01  
TCGA-A5-A0G9...-11D-A042-01  
TCGA-A5-A0GA...-11D-A042-01  
TCGA-A5-A0GB...-11D-A042-01  
TCGA-A5-A0GE...-11D-A042-01  
TCGA-A5-A0GG...-11D-A120-01  
TCGA-A5-A0GH...-21D-A042-01  
TCGA-A5-A0GI...-11D-A042-01  
TCGA-A5-A0GJ...-11D-A042-01  
TCGA-A5-A0GM...-11D-A042-01  
TCGA-A5-A0GN...-11D-A042-01  
TCGA-A5-A0GP...-11D-A042-01  
TCGA-A5-A0GQ...-11D-A120-01  
TCGA-A5-A0GR...-11D-A120-01  
TCGA-A5-A0GU...-11D-A042-01  
TCGA-A5-A0GV...-31D-A042-01  
TCGA-A5-A0G...11D-A042-01  
TCGA-A5-A0GX...-11D-A042-01  
TCGA-A5-A0R6...-11D-A102-01  
TCGA-A5-A0R7...-31D-A102-01  
TCGA-A5-A0R8...-11D-A102-01  
TCGA-A5-A0R9...-11D-A102-01  
TCGA-A5-A0RA...-21D-A102-01  
TCGA-A5-A0VP...-21D-A102-01

RefSeq genes

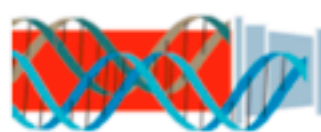
Available Datasets

- Available Datasets
  - Annotations (hg18)
  - The Cancer Genome Atlas (hg18) ⓘ
    - GBM Subtypes (Verhaak et. al.) ⓘ
    - GBM
    - Broad Firehose Standard Data Run: 2012\_03\_21 ⓘ
      - CESC
      - LUAD
      - UCEC
        - CopyNumber: [ genome\_wide\_snp\_6\_\_broad ]
        - Expression: [ agilentg4502a\_07\_3 ]
        - Methylation: [ humanmethylation27\_\_jhu\_usc ]
      - LUSC
      - KIRP
      - GBM
      - PANCANCER
      - LGG
      - LAML
      - BRCA
      - STAD
      - HNSC
      - PRAD

130M of 253M

Cancel OK

# AND EASY TO FIND & DOWNLOAD



**FIREHOSE**  
Broad GDAC

Home

Tools ▾

8 Added by [Aaron Ball](#), last edited by [Michael Noble](#) on Apr 21, 2012 ([view change](#))

- [AWG Reps](#)
- [Contact Us](#)
- [Dashboard-Analyses](#)
- [Dashboard-Stddata](#)
- [Data Usage Policy](#)
- [DCC Interactions](#)
- [FAQ](#)
- [Nozzle](#)
- [Pipeline Docs](#)
- [Presentations](#)
- [ProcessFlow](#)
- [QualityControl](#)

[Email Archive](#)

[Tracking System](#)

Summary of Data Status Data  
Reported into Broad GDAC Pipelines  
2012\_03\_21 10:00 AM PST

Category	Count
AnalysisReport	1000
ReleaseNotes	1000
Stddata	1000
Download	1000
Protected	1000
Open	1000
Success	1000
Failed	1000
Cancelled	1000
Skipped	1000
Ignored	1000
Other	1000

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\_03\_21 stddata Run

ReleaseNotes	# Datasets	% Processed	Download
<a href="#">BLCA</a>	8	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">BRCA</a>	16	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">CESC</a>	7	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">COADREAD</a>	14	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">GBM</a>	21	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">HNSC</a>	12	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRC</a>	16	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRP</a>	12	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LAML</a>	7	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LGG</a>	8	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LHC</a>	8	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUAD</a>	16	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUSC</a>	25	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">OV</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PAAD</a>	3	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PRAD</a>	5	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">SKCM</a>	1	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">STAD</a>	14	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">THCA</a>	7	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">UCEC</a>	16	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PANCANCER</a>	34	85%	<a href="#">Open</a> <a href="#">Protected</a>

2012\_03\_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
<a href="#">BRCA</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">COADREAD</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">GBM</a>	21	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LGG</a>	14	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUSC</a>	23	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">OV</a>	24	100%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRC</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LUAD</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">UCEC</a>	22	96%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">STAD</a>	14	93%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">KIRP</a>	16	89%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PRAD</a>	7	88%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">THCA</a>	7	88%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LAML</a>	11	85%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">BLCA</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">HNSC</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">LHC</a>	7	78%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">CESC</a>	6	60%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PAAD</a>	3	60%	<a href="#">Open</a> <a href="#">Protected</a>
<a href="#">PANCANCER</a>	11	58%	<a href="#">Open</a> <a href="#">Protected</a>

NEW: [firehose\\_get](#) utility to simplify retrieval of public result archives.

# AND EASY TO FIND & DOWNLOAD

## firehose\_get v0.3.0 (alpha)

---

Retrieving or utilizing TCGA results need not be difficult, especially for open-access data. To help simplify, we're currently alpha-testing the *firehose\_get* retrieval script. To join our alpha testing, simply [download the zip file from here](#), perform these 2 steps from a Unix-compatible command line

```
unix% unzip firehose_get.zip
unix% ./firehose_get
```

and follow the instructions shown below. Please note that downloading data from our site constitutes agreement to [this data usage policy](#).

```
firehose_get : retrieve open-access results of Broad Institute TCGA GDAC runs
Version: 0.3.0 alpha (Author: Michael S. Noble)

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

Two arguments are required; the first must be one of

    analyses | data | stddata

(the latter two values are equivalent), 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 the following
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

Flags:

-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
```



# 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

Aaron McKenna

Sachet Shukla

Raktim Sinha

Andrey Sivachenko

Carrie Sougnez

Petar Stojanov

Lihua Zhou

Hailei Zhang

Robert Zupko

## **Belfer-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



# Poster Contributions

- Poster : *Engineering Firehose* (DiCara et al)  
Poster : *RNA-Seq in Firehose* (Zhang et al)  
Poster : *GDAC Interoperability* (Cerami et al)  
Poster : *Broad SNP6 Pipeline* (Saksena et al)

THANK YOU!

<http://gdac.broadinstitute.org>

