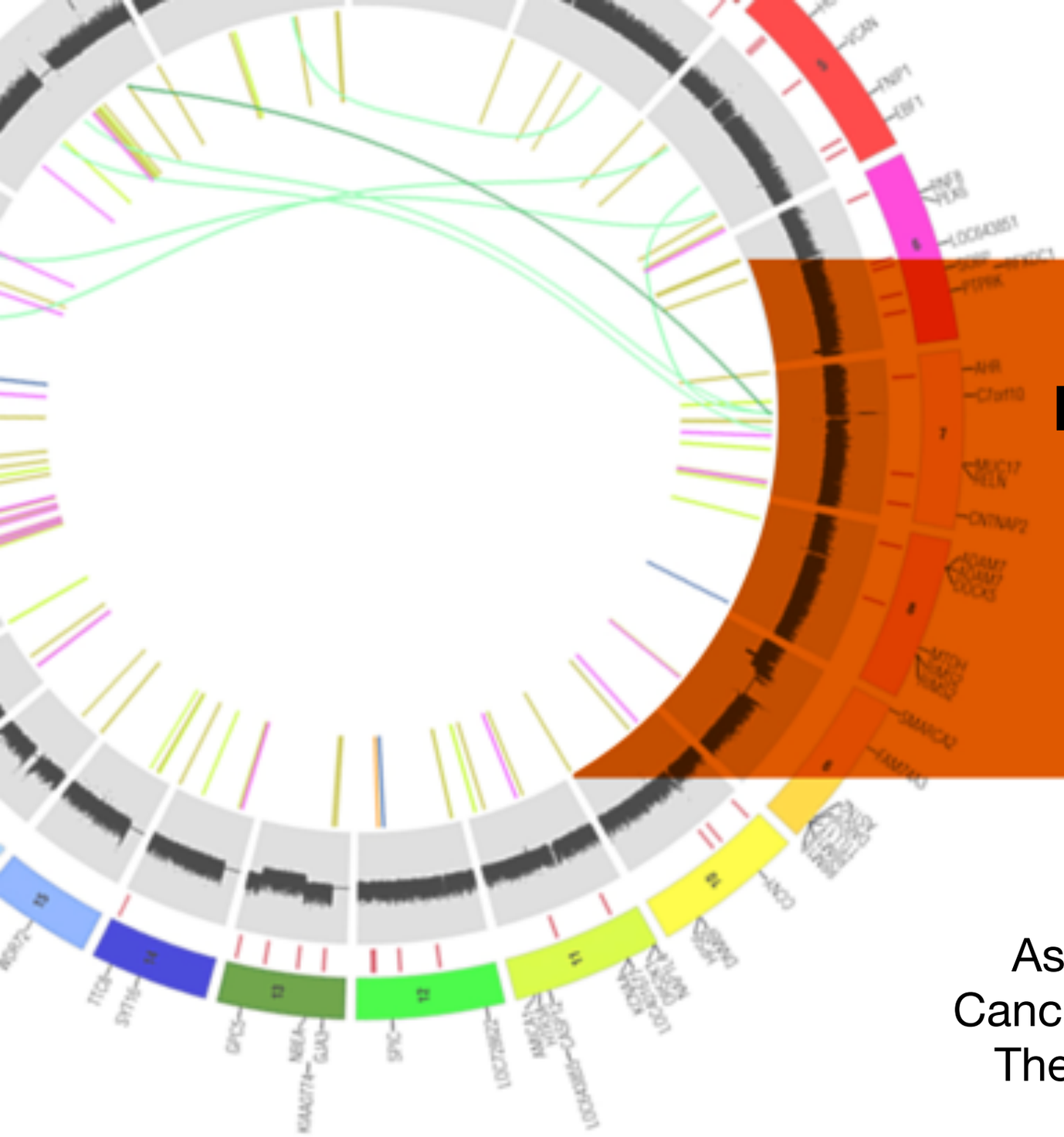


The Cancer Genome Atlas

TCGA Data & Analyses Beyond the DCC: Firehose

3rd TCGA Symposium
May 12, 2014
National Institutes of Health
Bethesda, MD

Michael S. Noble
Assistant Director for Data Science
Cancer Genome Computational Analysis
The Broad Institute of MIT & Harvard

Firehose Pipeline Manager
TCGA Genome Data Analysis Center

BROAD INSTITUTE

FIREHOSE
Broad GDAC

Acknowledgements

PI: Lynda Chin, Gaddy Getz

Broad Institute
Daniel DiCara
David Heiman
Harindra Arachchi
Hailei Zhang
Juok Cho
Jaegil Kim
Gordon Saksena
Douglas Voet
William Mallard
Michael Lawrence
Petar Stojanov
Lihua Zou
Chip Stewart
Scott Frazer
Pei Lin
Kristian Cibulskis
Lee Lichtenstein
Aaron McKenna
Andrey Sivachenko
Carrie Sougnez
Lee Lichtenstein
Steven Schumacher
Raktim Sinha

Belfer/DFCI/MDACC
Juinhua Zhang
Spring Liu
Sachet Shukla
Terrence Wu

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

Broad Institute Leadership: Todd Golub, Eric Lander

Harvard Medical School

Matthew Meyerson
Andrew Cherniack
Juliann Chmielecki
Rameen Beroukhim
Scott Carter

Peter Park
Nils Gehlenborg
Semin Lee
Richard Park

In Particular

Daniel DiCara
David Heiman
Harindra Arachchi
Hailei Zhang
Juok Cho
Jaegil Kim

who have worked tirelessly over the past
2-3 years, SIMULTANEOUSLY creating, extending, &
supporting AWG & standard runs, methods, pipelines,
contributing research results to papers, developing
infrastructure, curating data & mining Firehose results

Begin

Did you know you can obtain comprehensive genomic profiles of 30 cancer cohorts, across 60K sample aliquots, with a single command?

Did you know you can obtain
comprehensive genomic profiles of
30 cancer cohorts, across 60K sample
aliquots, with a single command?

```
linux%   firehose_get   analyses  latest
```

~1100 analysis result pkgs in *single run* *(May 2014)*
988 Nozzle HTML reports
Each citable in literature via DOIs

Our GDAC distills ~40 TB input data down to 9GB results
3X orders of magnitude

Why?

Because The Bad Old Days ...

Of solitary, manual experimentation on small sample sets ...

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

Then get distracted, forget ...
Search, run again, ... lose track, search ...
Repeat ... for 20 more disease types
GBM, LUNG, AML, ...

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

Don't Scale to TCGA

GDAC Firehose data stream

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	108	99	0	138	0	96	0	124	54	28
BRCA	914	866	874	0	889	529	805	0	868	408	507
CESC	122	32	102	0	122	0	0	0	122	0	36
COAD	423	423	413	69	420	155	192	0	407	269	155
COADREAD	592	591	575	104	582	224	264	0	550	399	224
DLBC	28	0	17	0	17	0	0	0	16	0	0
GBM	598	565	563	0	411	542	161	491	0	214	276
HNSC	328									212	0
KICH	66									0	0
KIRC	502									454	403
KIRP	149									0	0
LAML	202									0	199
LGG	222									0	0
LIHC	99									0	0
LUAD	439									237	229
LUSC	376									195	178
OV	592									412	316
PAAD	57									0	0
PANCAN8	4086	3882	3907	210	3798	2150	2515	1061	3169	2282	2152
PRAD	180	127	171	0	172	0	140	0	170	0	83
READ	169	168	162	35	162	69	72	0	143	130	69
SARC	29	0	29	0	29	0	0	0	29	0	0
SKCM	273	138	253	101	253	0	247	0	240	164	0
STAD	238	162	144	0	145	0	43	0	134	0	116
THCA	435	218	330	94	353	0	254	0	349	224	323
UCEC	512	451	493	106	500	54	333	0	485	200	248
Totals	7106	5839	6195	501	6443	2225	4357	1061	5627	3173	3166
	+1830	+1665	+2021	+501	+4181		+4357		+5267	+3173	+1142

30x11x181 dimensional file space

>60K sample aliquots today

2011-2012: ~24K new in a single year

Nothing like this had ever been attempted

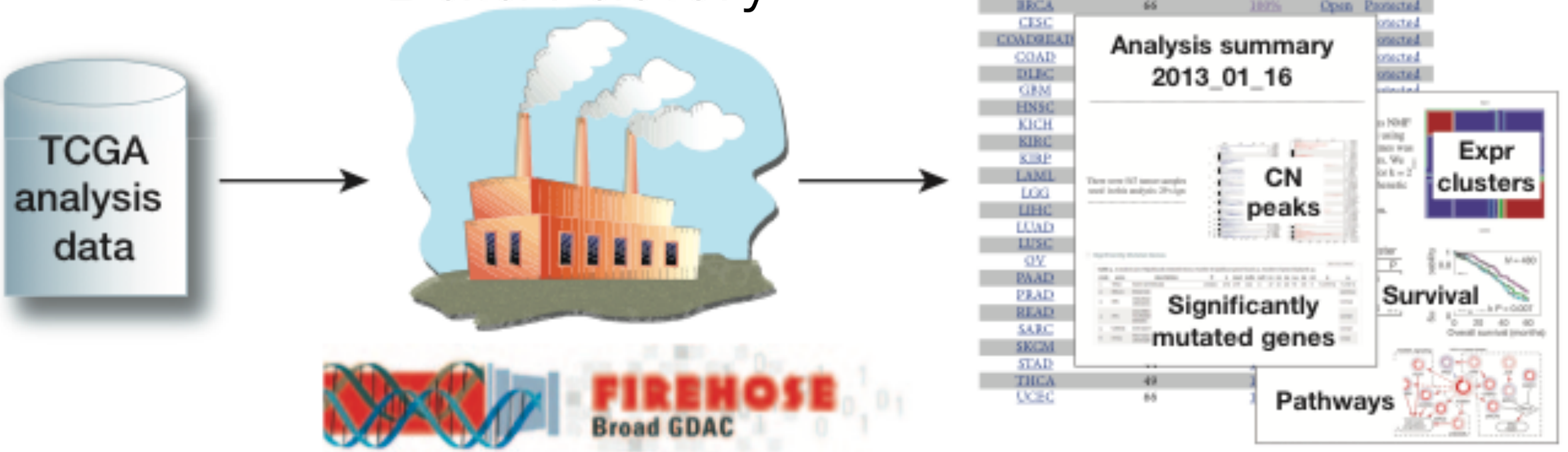
Context : 2-3 orders magnitude shift

Exome Sequencing Studies of Cancer in 2011			
Cancer Type	**#Samples**	**Key Finding(s)**	**Publication**
Melanoma	14 cases	Frequent mutations in GRIN2A	Wei et al. Nat Genet. 2011.
Metastatic Melanoma	8 cell lines	Mutations in MAP3K5 and MAP3K9	Stark et al. Nat Genet. 2011
Melanoma	7 cell lines	Recurring somatic MAP2K1 and MAP2K2 mutations (8%)	Nikolaev et al. Nat Genet. 2011
Head and neck squamous cell	74 cases	Mutations in TP53, CDKN2A, PIK3CA, HRAS, and squamous differentiation genes.	Stransky et al. Science.
Head and neck squamous cell	32 cases	Mutations in TP53, CDKN2A, PIK3CA, and HRAS, FBXW7 and NOTCH1. Tumor-suppressor role for NOTCH1.	Agrawal et al. Science 2011.
Renal carcinoma	7 cases	Frequent mutation of the SWI/SNF complex gene PBRM1	Varela et al. Nature 2011.
Pancreatic cancer	15 cell lines	Genomic instability caused by MLH1 haploinsufficiency and complete deficiency	Wang et al. Genome Res. 2011
Pancreatic neoplastic cysts	8 cyst resections	Recurrent mutations in components of ubiquitin-dependent pathways	Wu et al. PNAS 2011.
Gastric cancer	22 cases	Frequent mutation of ARID1A	Wang et al. Nat Genet 2011.
Prostate cancer	3 primaries 16 metastases	Recurrent alterations in TP53, DLK2, GPC6, and SDF4	Kumar et al. PNAS 2011

http://massgenomics.org/2012/01/cancer-genome-and-exome-sequencing-in-2011.html

Acute Need for Automation, Systematic Rigor, and Transparency

Data Factory



2,~~5~~00+ pipelines per month, across all TCGA disease types

Results dashboards and biologist-friendly reports

Open to public for browsing and automatic download

Democratize TCGA science by lowering entry barriers

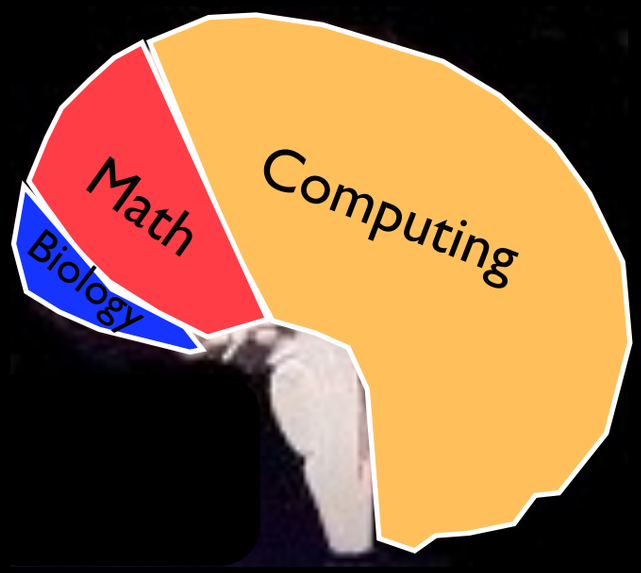
6000

Drilling into big cancer-genome data, ***Nature Methods*** **10**, 293–297 (2013)

But as clear, simple, accessible as possible

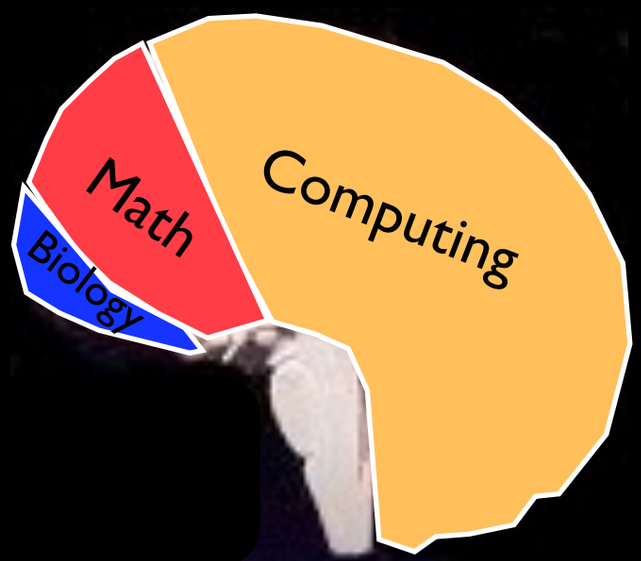
This is Your
Researcher
Brain

But as clear, simple, accessible as possible

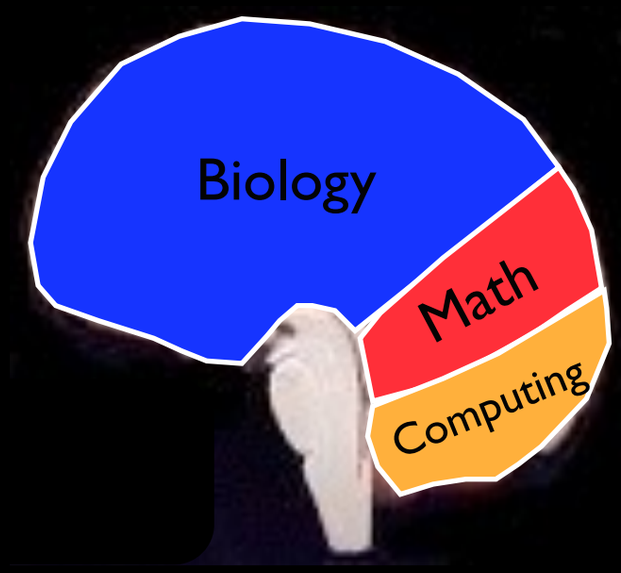


When Coding
Or Data
Exploration
Is Hard

But as clear, simple, accessible as possible



When Coding
Or Data
Exploration
Is Hard



When
Easier

Easy stuff should stay easy, so hard stuff becomes possible

So Our GDAC Firehose Generates

1

Version-stamped, standardized datasets

- Precursor to automated analyses: aggregates all available sample batches
- Into a single, uniformly-formatted bolus (one per disease X datatype), which can be
- Immediately fed to algorithmic codes without further data preparation
- **Monthly**

2

Version-stamped package of standard analyses results

- Automatically generated for dozens of algorithms: GISTIC, MutSig, Clustering, Correlation, …
- **Quarterly**

3

Version-stamped, biologist-friendly reports

- Encapsulating analysis results in a form accessible to a wide audience
- Online for public browsing
- ***Citable in the literature through DOIs***

So Our GDAC Firehose Generates

1 **Version-stamped**, standardized datasets

- Precursor to automated analyses: aggregates all available sample batches
- Into a single, uniformly-formatted bolus (one per disease X datatype), which can be
- Immediately fed to algorithmic codes without further data preparation
- **Monthly**

2 **Version-stamped** package of standard analyses results

- Automatically generated for dozens of algorithms: GISTIC, MutSig, Clustering, Correlation, …
- **Quarterly**

3 **Version-stamped**, biologist-friendly reports

- Encapsulating analysis results in a form accessible to a wide audience
- Online for public browsing
- ***Citable in the literature through DOIs***

Rigorous Data Science

↓

Credible Biology

And More Recently …

4 Custom runs tailored to TCGA AWGs

Currency: pipelines can be run on the *latest snapshot of data* from DCC, *avoiding the time & sample lag of monthly runs*

Flexibility: easily include AWG-curated disease subtypes, even custom analyses

Speed: usually executed in only a few days time

Familiarity: using same internal Firehose machinery, external-facing dashboards, Nozzle, `firehose_get` etc known to community

39 AWG runs performed in 2013
Plus 23 standard data & analyses runs
>5 runs per month in 2013

Results Couched in biologist-friendly online reports

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 ≤ 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.

Organized like a paper

- Overview (“Abstract”)
- Results (with download link)
- Methods
- References

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 ≤ 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.

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	1.81476-87	8.49459e-76	chr19:34947990-35023082	1
3q26.2	1.07226-60	1.07226-60	chr3:170905217-170925258	0 [MECOM]

Ovarian Serous Cystadenocarcinoma: Clustering of mRNA expression: consensus NMF

Maintained by Robert Zapko *(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.

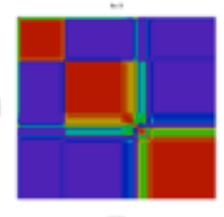
Results

Gene expression patterns of molecular subtypes

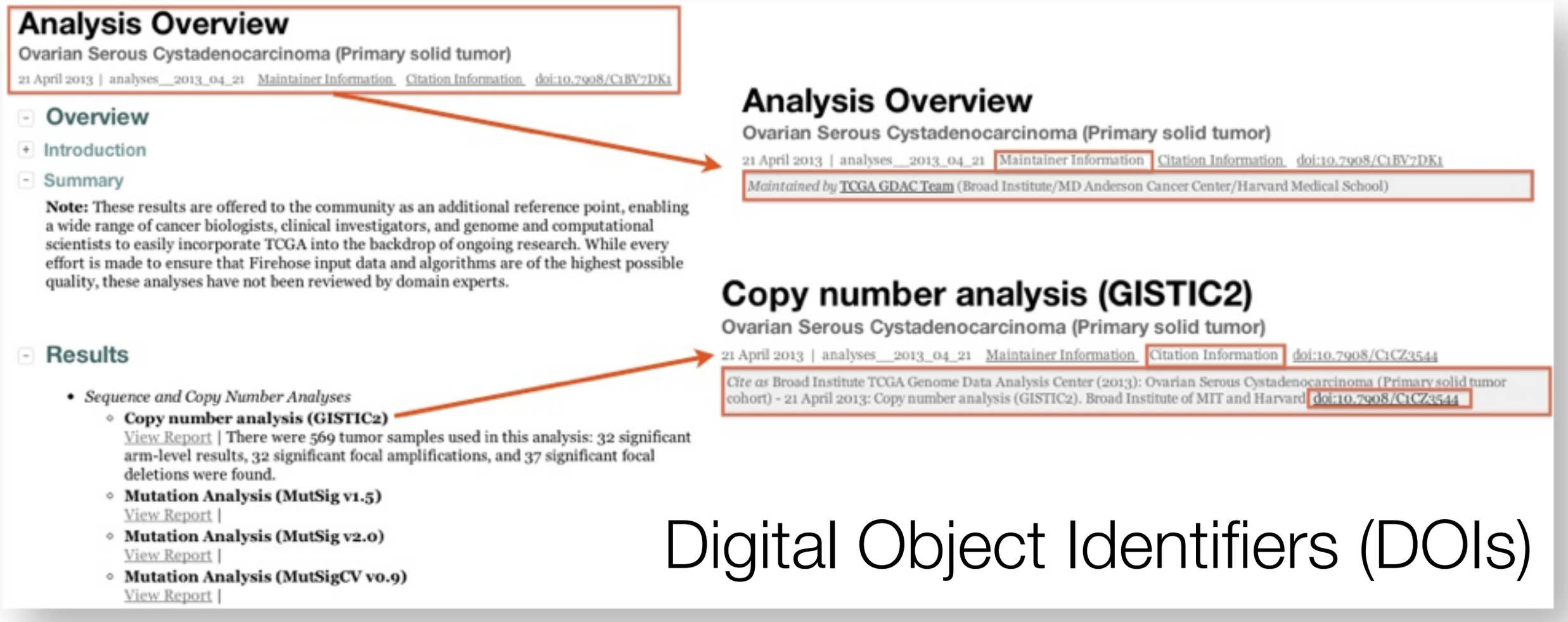
Consensus and correlation matrix

GET HIGH-RES IMAGE

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



Directly Citable in The Literature



Digital Object Identifiers (DOIs)

~ 1,000 reports generated per analysis run, thousands of pages of results
First of its kind at Broad Institute: nothing at this scale, anywhere?

With Dead Simple Bulk Retrieval

```
firehose_get : retrieve open-access results of Broad Institute TCGA GDAC runs
Version: 0.3.3 (Author: Michael S. Noble)

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

firehose_get

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

Simple 20K bash script, just 1 moving part

Or, if you prefer
interactive browsing

BROAD INSTITUTE

FIREHOSE
Broad GDAC

Liberating scientific knowledge from data, at scale

About | Dashboards | Data | Analyses | Software | Documentation | FAQ | Download | Contact Us | What's New?

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Download
Bladder urothelial carcinoma	BLCA	311	Browse	Download
Breast invasive carcinoma	BRCA	1061	Browse	Download
Cervical and endocervical cancers	CESC	257	Browse	Download
Colon adenocarcinoma	COAD	448	Browse	Download
Colorectal adenocarcinoma	COADREAD	616	Browse	Download
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	53	Browse	Download
Esophageal carcinoma	ESCA	176	Browse	Download
Glioblastoma multiforme	GBM	607	Browse	Download
Head and Neck squamous cell carcinoma	HNSC	517	Browse	Download
Kidney Chromophobe	KICH	113	Browse	Download
Kidney renal clear cell carcinoma	KIRC	536	Browse	Download
Kidney renal papillary cell carcinoma	KIRP	274	Browse	Download
Acute Myeloid Leukemia	LAML	200	Browse	Download
Brain Lower Grade Glioma	LGG	516	Browse	Download
Liver hepatocellular carcinoma	LIHC	273	Browse	Download
Lung adenocarcinoma	LUAD	563	Browse	Download
Lung squamous cell carcinoma	LUSC	493	Browse	Download
Mesothelioma	MESO	37	Browse	Download
Ovarian serous cystadenocarcinoma	OV	592	Browse	Download
Pancreatic adenocarcinoma	PAAD	131	Browse	Download
Pheochromocytoma and Paraganglioma	PCPG	179	Browse	Download
Prostate adenocarcinoma	PRAD	427	Browse	Download
Rectum adenocarcinoma	READ	168	Browse	Download
Sarcoma	SARC	217	Browse	Download
Skin Cutaneous Melanoma	SKCM	448	Browse	Download
Stomach adenocarcinoma	STAD	373	Browse	Download
Thyroid carcinoma	THCA	496	Browse	Download
Uterine Corpus Endometrial Carcinoma	UCEC	556	Browse	Download
Uterine Carcinosarcoma	UCS	57	Browse	Download

Starting Point For Most GDAC Questions

Open-Source Look/Feel

FAQ
Release Notes

Searchable Mail Archive

Search Results

gdac.broadinstitute.org

Analyses?

Analysis Overview for OV-TP

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

Results

- *Sequence and Copy Number Analyses*
  - **Copy number analysis (GISTIC**
    View Report | There were 559 tumo
    focal amplifications, and 39 signific
  - **Mutation Analysis (MutSig v2.**
    View Report |
  - **Mutation Analysis (MutSig vS2**
    View Report |
- *Clustering Analyses*
  - **Clustering of copy number data: consensus NMF**
    View Report | The most robust consensus NMF clustering of *559* samples using the *70* copy number focal regions 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 Methylation: consensus NMF**
    View Report | The *2363* 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,

Crown Jewels
GISTIC & MutSig
(CopyNumber & Mutation significance)

Clusterings for Most Datatypes
mRNA, miR, *-Seq, RPPA
CopyNumber, Methylation (27 & 450)

412 samples and *150* proteins identified *4* 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 mRNA expression: consensus NMF**
  View Report | The most robust consensus NMF clustering of *569* 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 *569* 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 mRNAseq gene expression: consensus NMF**
  View Report | The most robust consensus NMF clustering of *262* 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.

- **PARADIGM pathway analysis of mRNA expression data**
  View Report | There were 62 significant pathways identified in this analysis.
- **PARADIGM pathway analysis of mRNA expression and copy number data**
  View Report | There were 76 significant pathways identified in this analysis.
- **PARADIGM pathway analysis of mRNASeq expression data**

Pathway
Paradigm (Stuart et al, UCSC)
HotNet (Raphael et al, Brown)

Correlations : 19 currently available
vs clinical (most important)
vs clusters,
datatype vs. datatype

- *Correlation Analyses*
  - **Correlation between copy number variati**
    View Report | Testing the association between c
    across 552 patients, 8 significant findings detect
  - **Correlation between copy number variati**
    View Report | Testing the association between c
    across 552 patients, 12 significant findings detec
  - **Correlation between gene methylation st**
    View Report | Testing the association between 1
    thresholded by Q value < 0.05, 1 clinical feature
  - **Correlation between molecular cancer subtypes and selected clinical features**
    View Report | Testing the association between subtypes identified by 12 different clustering approaches and 6 clinical features across 578 patients, 13 significant findings detected with P value < 0.05.
  - **Correlations between copy number and mRNAseq expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are 1087.4, 1797, 2427, 3136.6, 3915, 4708, 5472.8, 6145.2, 6816, respectively.
  - **Correlations between copy number and miR expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are -0.03696, -0.01514, -5e-04, 0.0203, 0.0452, 0.09412, 0.1859, 0.27658, 0.37064, respectively.
  - **Correlation between mRNA expression and DNA methylation**
    View Report | The top 25 correlated methylation probes per gene are displayed. Total number of matched samples = 262. Number of gene expression samples = 262. Number of methylation samples = 262.

- **PARADIGM pathway analysis of mRNA expression data**
  View Report | There were 62 significant pathways identified in this analysis.
- **PARADIGM pathway analysis of mRNA expression and copy number data**
  View Report | There were 76 significant pathways identified in this analysis.
- **PARADIGM pathway analysis of mRNASeq expression data**

Pathway
Paradigm (Stuart et al, UCSC)
HotNet (Raphael et al, Brown)

- *Correlation Analyses*
  - **Correlation between copy number variati**
    View Report | Testing the association between c
    across 552 patients, 8 significant findings detect
  - **Correlation between copy number variati**
    View Report | Testing the association between c
    across 552 patients, 12 significant findings detec
  - **Correlation between gene methylation st**
    View Report | Testing the association between 1
    thresholded by Q value < 0.05, 1 clinical feature
  - **Correlation between molecular cancer subtypes and selected clinical features**
    View Report | Testing the association between subtypes identified by 12 different clustering approaches and 6 clinical features across 578 patients, 13 significant findings detected with P value < 0.05.
  - **Correlations between copy number and mRNAseq expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are 1087.4, 1797, 2427, 3136.6, 3915, 4708, 5472.8, 6145.2, 6816, respectively.
  - **Correlations between copy number and miR expression**
    View Report | The correlation coefficients in 10, 20, 30, 40, 50, 60, 70, 80, 90 percentiles are -0.03696, -0.01514, -5e-04, 0.0203, 0.0452, 0.09412, 0.1859, 0.27658, 0.37064, respectively.
  - **Correlation between mRNA expression and DNA methylation**
    View Report | The top 25 correlated methylation probes per gene are displayed. Total number of matched samples = 262. Number of gene expression samples = 262. Number of methylation samples = 262.

Correlations : 19 currently available
vs clinical (most important)
vs clusters,
datatype vs. datatype
even custom data ...

Impact

Established Traction as Nexus Resource

	Pages	Hits	Bandwidth
Interactive Use	**643,858** (221.18 Pages/Visit)	**757,376** (260.17 Hits/Visit)	**277.61 GB** (99997.5 KB/Visit)
firehose_get downloads		108,397+1198	1567.50 GB

May 2013 640K pages 860K hits 1.8 TB traffic

July 2013	**> 2**	**TB traffic**
April 2014	**~ 6**	**TB traffic**

- Across dozens of centers & portals
- Research / Academic / Commercial
- International scope

Democratize TCGA science: lower entry barriers

"Oh, that's interesting, maybe my code has found something here ... I wonder if this is seen in the Firehose version 2013_04_21 results, too?"

It's like a free expert assistant / second opinion

Extremely low hanging fruit!

At TCGA scale one might otherwise need to ...

Spend weeks/months obtaining protected data credentials

Or hire more staff

Or become a TCGA data guru, obtaining samples spread across many files

Then more time, mastering the analytics

Complexity & volume preclude this approach for many individuals

Automated, High-Throughput Clinical Miner

Firehose automatically mines entire suite of clinical params to identify statistically significant relationships with every TCGA datatype (e.g. SMGs) or aggregate (e.g. clusters)

The results, which e.g. include survival curves (when possible) for every TCGA disease, are posted openly on the Broad

Since automation is “free,” these don’t have to be 100% to establish potentially interesting signposts

Clinical Correlations vs Clusters

Clinical Features	Statistical Tests	Copy Number Ratio CNMF subtypes	METHLYATION CNMF	RPPA CNMF subtypes	RPPA cHierClus subtypes	RNAseq CNMF subtypes	RNAseq cHierClus subtypes	MIRSEQ CNMF	MIRSEQ CHIERARCHICAL	MIRseq Mature CNMF subtypes	MIRseq Mature cHierClus subtypes
Time to Death	logrank test	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)	100 (1.00)
AGE	ANOVA	0.111 (1.00)	0.00114 (0.176)	0.0268 (1.00)	0.0567 (1.00)	0.585 (1.00)	0.386 (1.00)	0.733 (1.00)	0.667 (1.00)	0.356 (1.00)	0.398 (1.00)
PATHOLOGY T STAGE	Chi-square test	0.000171 (0.0275)	0.0519 (1.00)	0.0267 (1.00)	0.0581 (1.00)	0.43 (1.00)	0.929 (1.00)	0.11 (1.00)	0.000724 (0.114)	0.0866 (1.00)	0.0914 (1.00)
PATHOLOGY N STAGE	Fisher's exact test	5.97e-05 (0.00973)	0.0326 (1.00)	0.031 (1.00)	0.0397 (1.00)	0.0228 (1.00)	0.162 (1.00)	0.163 (1.00)	0.164 (1.00)	0.111 (1.00)	0.111 (1.00)
COMPLETENESS OF RESECTION	Chi-square test	0.224 (1.00)	0.306 (1.00)	0.0798 (1.00)	0.0217 (1.00)	0.203 (1.00)	0.0353 (1.00)	0.187 (1.00)	0.478 (1.00)	0.229 (1.00)	0.198 (1.00)
NUMBER OF LYMPH NODES	ANOVA	0.00012 (0.0194)	0.0477 (1.00)	0.0366 (1.00)	0.0285 (1.00)	0.0959 (1.00)	0.166 (1.00)	0.11 (1.00)	0.0746 (1.00)	0.0798 (1.00)	0.0798 (1.00)
GLEASON SCORE COMBINED	ANOVA	8.19e-07 (0.000137)	0.0113 (1.00)	0.00449 (0.651)	0.00912 (1.00)	0.286 (1.00)	0.107 (1.00)	0.187 (1.00)	0.336 (1.00)	0.372 (1.00)	0.376 (1.00)
GLEASON SCORE PRIMARY	ANOVA	8.24e-07 (0.000137)	0.00669 (0.943)	0.000644 (0.102)	0.000586 (0.0938)	0.0111 (1.00)	0.00145 (0.217)	0.0679 (1.00)	0.632 (1.00)	0.611 (1.00)	0.896 (1.00)
GLEASON SCORE SECONDARY	ANOVA	0.253 (1.00)	0.722 (1.00)	0.693 (1.00)	0.573 (1.00)	0.397 (1.00)	0.542 (1.00)	0.0917 (1.00)	0.347 (1.00)	0.512 (1.00)	0.422 (1.00)
GLEASON SCORE	ANOVA	6.03e-08 (1.01e-05)	0.00601 (0.854)	0.00141 (0.215)	0.00143 (0.216)	0.172 (1.00)	0.0518 (1.00)	0.115 (1.00)	0.54 (1.00)	0.193 (1.00)	0.191 (1.00)
PSA RESULT PREOP	ANOVA	0.0489 (1.00)	0.000992 (0.155)	0.0347 (1.00)	0.248 (1.00)	0.0028 (0.418)	0.0547 (1.00)	0.0969 (1.00)	0.167 (1.00)	0.0687 (1.00)	0.0621 (1.00)
DAYS TO PREOP PSA	ANOVA	0.689 (1.00)	0.588 (1.00)	0.00116 (0.178)	0.00137 (0.21)	0.879 (1.00)	0.561 (1.00)	0.086 (1.00)	0.0187 (1.00)	0.0103 (1.00)	0.00805 (1.00)
PSA VALUE	ANOVA	0.148 (1.00)	0.0822 (1.00)	0.18 (1.00)	0.409 (1.00)	0.302 (1.00)	0.00387 (0.569)	0.021 (1.00)	0.0395 (1.00)	0.0392 (1.00)	0.0477 (1.00)
DAYS TO PSA	ANOVA	0.88 (1.00)	0.128 (1.00)	0.256 (1.00)	0.0928 (1.00)	0.0337 (1.00)	0.411 (1.00)	0.633 (1.00)	0.34 (1.00)	0.156 (1.00)	0.224 (1.00)
CURATED FINAL CELLULARITY	Chi-square test	0.126 (1.00)	0.01 (1.00)	0.00917 (1.00)	0.00392 (0.572)	0.0045 (0.651)	0.0129 (1.00)	0.102 (1.00)	0.0195 (1.00)	0.0295 (1.00)	0.0715 (1.00)
CURATED FINAL GLEASON	Chi-square test	6.57e-09 (1.11e-06)	0.0234 (1.00)	0.00334 (0.494)	0.0274 (1.00)	0.079 (1.00)	0.0237 (1.00)	0.252 (1.00)	0.484 (1.00)	0.197 (1.00)	0.131 (1.00)
CURATED TOTAL FINAL GLEASON	ANOVA	7.57e-11 (1.29e-08)	0.000857 (0.135)	2.67e-06 (0.000441)	5e-05 (0.0082)	0.00592 (0.846)	0.0112 (1.00)	0.0859 (1.00)	0.473 (1.00)	0.68 (1.00)	0.442 (1.00)

http://gdac.broadinstitute.org/runs/awg_prad__2014_03_14/reports/cancer/PRAD-TP/Correlate_Clinical_vs_Molecular_Subtypes/nozzle.html

PRAD awg run (D. Heiman)

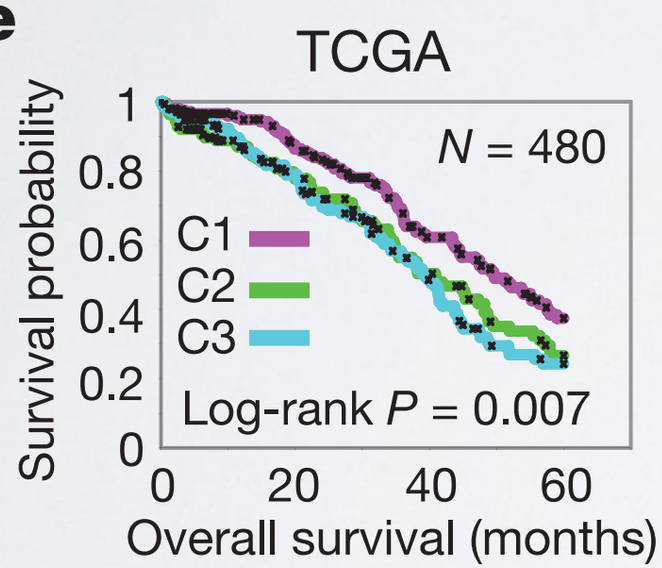
Novel discoveries lurk in Firehose outputs

d

		Gene cluster			
		D	I	M	P
miRNA cluster	C1	55	48	15	89
	C2	40	21	51	29
	C3	39	37	43	20

CNMF clustering of Ovarian miR expression yielded 3 subtypes

e



One of which correlated to significantly longer survivability

Integrated genomic analyses of ovarian carcinoma
TCGA Network, Nature, 2011

What's Next?

We’re not perfect, and we try hard to make good stuff

BUT WE MAKE A LOT

Not always easy to navigate, assimilate, and query

So, after having bred the beast
And making it easier to robustly feed, at scale

We now have time to make its volume & complexity even more accessible …

Unified Home Dashboard

Quickly browse to most important content

Disease Name	Cohort	Cases	Analyses	Archives
Adrenocortical carcinoma	ACC	92	Browse	Download
Bladder urothelial carcinoma	BLCA	311	Browse	Download
Breast invasive carcinoma	BRCA	1061	Browse	Download
Cervical and endocervical cancers	CESC	257	Browse	Download
Colon adenocarcinoma	COAD	448	Browse	Download
Colorectal adenocarcinoma	COADREAD	616	Browse	Download
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	53	Browse	Download
Esophageal carcinoma	ESCA	176	Browse	Download
Glioblastoma multiforme	GBM	607	Browse	Download
Head and Neck squamous cell carcinoma	HNSC	517	Browse	Download
Kidney Chromophobe	KICH	113	Browse	Download
Kidney renal clear cell carcinoma	KIRC	536	Browse	Download
Kidney renal papillary cell carcinoma	KIRP	274	Browse	Download
Acute Myeloid Leukemia	LAML	200	Browse	Download
Brain Lower Grade Glioma	LGG	516	Browse	Download
Liver hepatocellular carcinoma	LIHC	273	Browse	Download
Lung adenocarcinoma	LUAD	563	Browse	Download
Lung squamous cell carcinoma	LUSC	493	Browse	Download
Mesothelioma	MESO	37	None	Download
Ovarian serous cystadenocarcinoma	OV	592	Browse	Download
Pancreatic adenocarcinoma	PAAD	131	Browse	Download
Pheochromocytoma and Paraganglioma	PCPG	179	None	Download
Prostate adenocarcinoma	PRAD	427	Browse	Download
Rectum adenocarcinoma	READ	168	Browse	Download
Sarcoma	SARC	217	Browse	Download
Skin Cutaneous Melanoma	SKCM	448	Browse	Download
Stomach adenocarcinoma	STAD	373	Browse	Download
Thyroid carcinoma	THCA	496	Browse	Download

Data Summary for Breast invasive carcinoma

Cohort	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNASeq	miR	miRSeq	RPPA	MAF
BRCA	1061	988	1042	19	1044	526	1037	0	1031	408	976

Browse Samples | Browse Workflow Graph | Browse Analyses

Breast invasive carcinoma (BRCA) Samples Report

Overview

Introduction

Summary

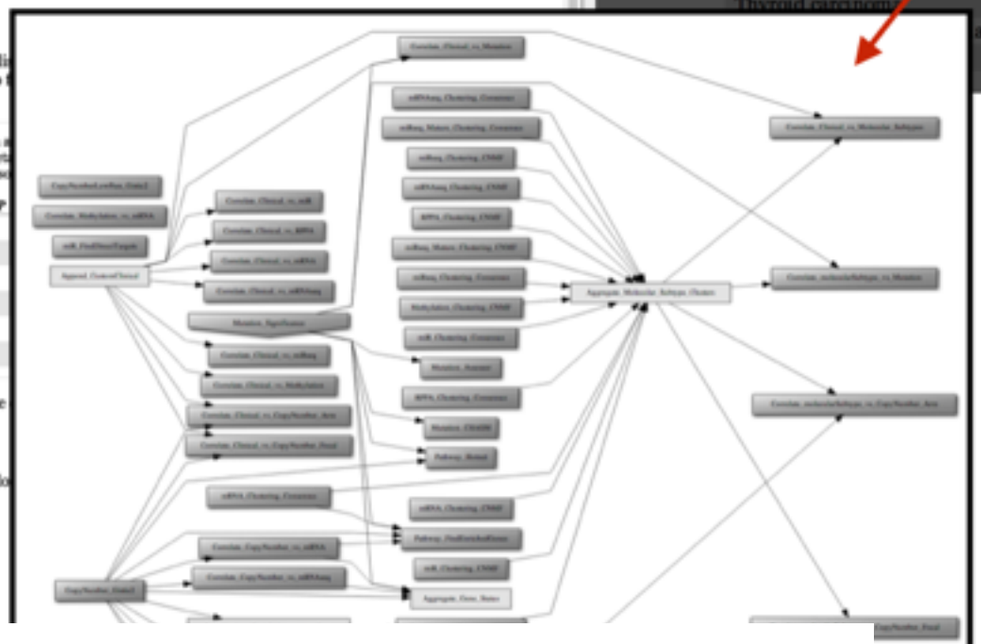
There were 15 redactions, 35 replicate aliquots, 11 blackli
sample counts for those samples that were ingested into
segregating FFPEs.

Table 1. This table provides a breakdown of sample counts on a
containing a list of the samples that comprise that count and det
Please note, there are usually multiple protocols per data type, s

Sample Type	BCR	Clinical	CN	LowP
TP	1061	988	1042	19
TM	7	6	7	0
NB	962	921	932	19
NT	139	154	131	0
FFPE	0	0	4	0
Totals	1061	988	1042	19

The sample type short letter codes in the table above are

- TP: Primary Solid Tumor
- TR: Recurrent Solid Tumor
- TB: Primary Blood Derived Cancer - Peripheral Blo
- TAP: Additional - New Primary
- TM: Metastatic
- TAM: Additional Metastatic
- NB: Blood Derived Normal
- NT: Solid Tissue Normal



Analysis Overview

Breast Invasive Carcinoma (Primary solid tumor)

16 April 2014 | analyses__2014_04_16 Maintainer Information Citation Information doi:10.7908/C1BR8Q8B

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*
  - **CHASM 1.0.5 (Cancer-Specific High-throughput Annotation of Somatic Mutations)**
    View Report | There are 49047 mutations identified by MuTect and 3933 mutations with significant functional impact at BHFDR <= 0.25.
  - **LowPass Copy number analysis (GISTIC2)**
    View Report | There were 19 tumor samples used in this analysis: 15 significant arm-level results, 2 significant focal amplifications, and 2 significant focal deletions were found.
  - **Mutation Analysis (MutSig v1.5)**
    View Report |
  - **Mutation Analysis (MutSig v2.0 and MutSigCV v0.9 merged result)**
    View Report |
  - **Mutation Analysis (MutSig v2.0)**
    View Report |
  - **Mutation Analysis (MutSigCV v0.9)**
    View Report |
  - **Mutation Assessor**
    View Report |
  - **SNP6 Copy number analysis (GISTIC2)**

Clickable workflow DAG

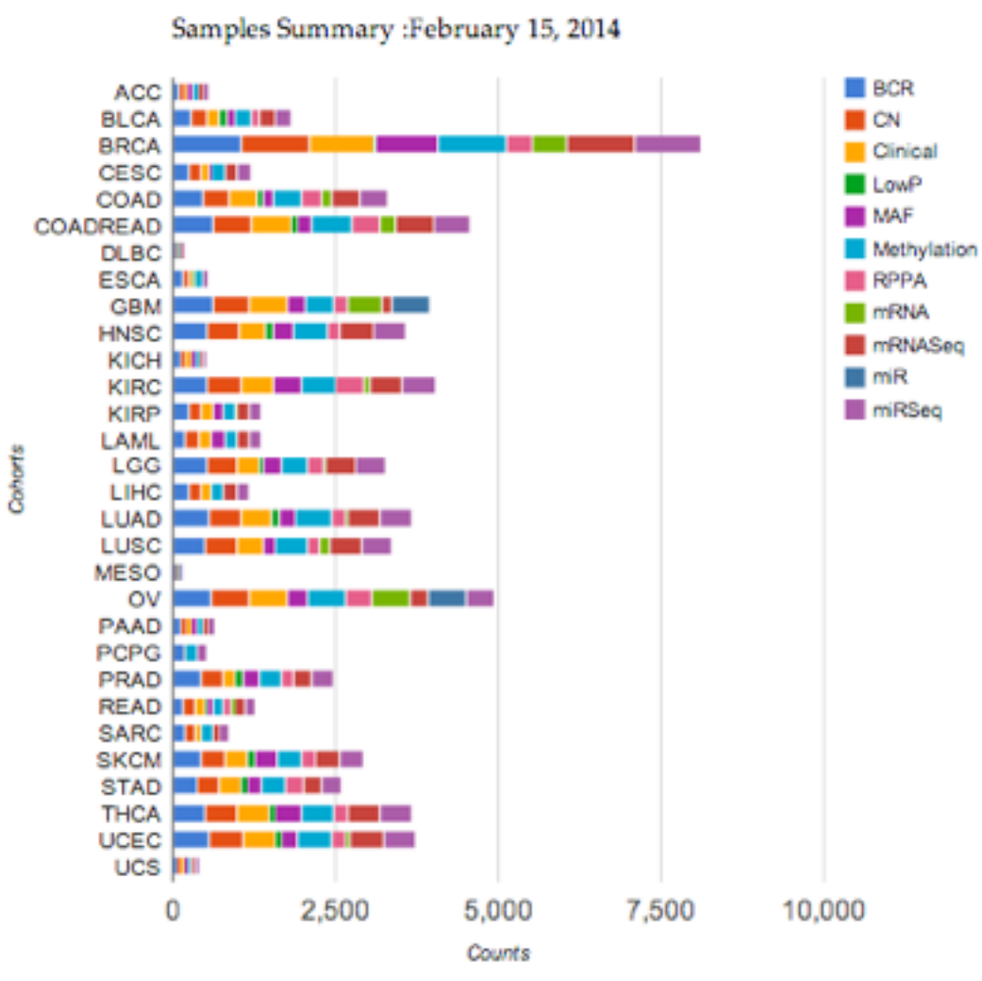
API-Powered Firehose Browser

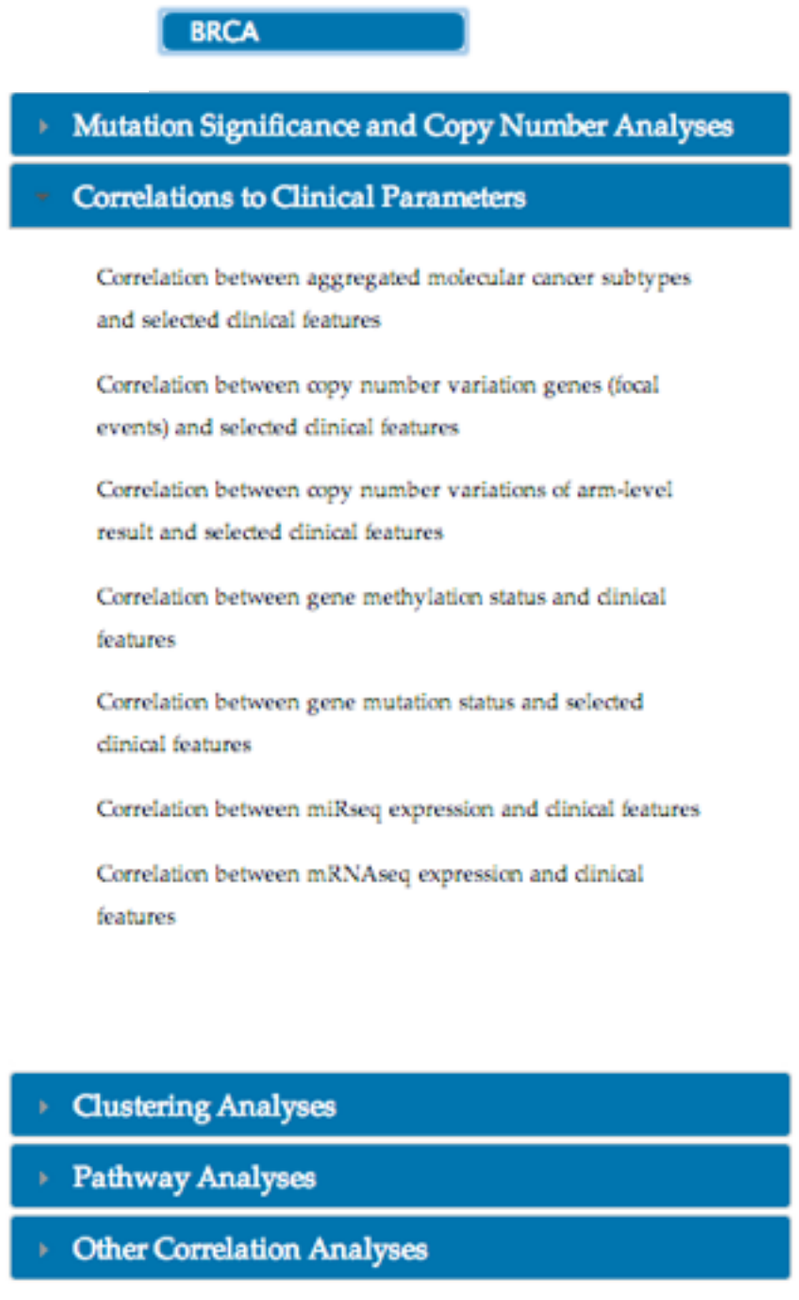
BROAD INSTITUTE

FIREHOSE Broad GDAC

Genome Data Analysis Center

Welcome to the interactive data portal of the Broad Institute Firehose Genome Data Analysis Center of The Cancer Genome Atlas.





H. Arachchi & Team

API-Powered Firehose Browser

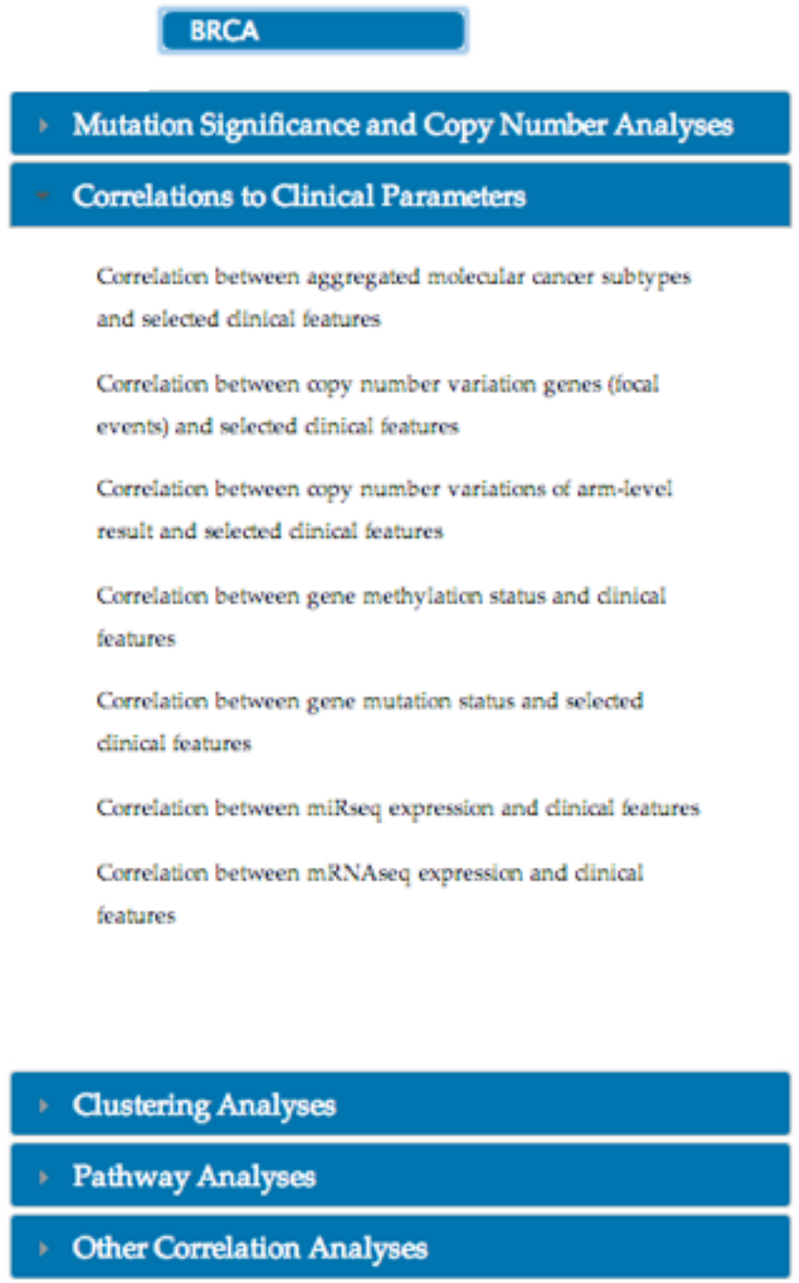
BROAD INSTITUTE

FIREHOSE Broad GDAC

Genome Data Analysis Center

Welcome to the interactive data portal of the Broad Institute Firehose Genome Data Analysis Center of The Cancer Genome Atlas.

~1000 Reports
Find Yours in 2 Clicks



H. Arachchi & Team

API-Powered Firehose Browser

BROAD INSTITUTE

FIREHOSE Broad GDAC

Genome Data Analysis Center

Welcome to the interactive data portal of the Broad Institute Firehose Genome Data Analysis Center of The Cancer Genome Atlas.

~1000 Reports
Find Yours in 2 Clicks

BRCA ← Cohort

- Mutation Significance and Copy Number Analyses
- Correlations to Clinical Parameters

Correlation between aggregated molecular cancer subtypes and selected clinical features

Correlation between copy number variation g events) and selected clinical features

Correlation between copy number variations of arm-level result and selected clinical features

Correlation between gene methylation status and clinical features

Correlation between gene mutation status and selected clinical features

Correlation between miRseq expression and clinical features

Correlation between mRNAseq expression and clinical features

DataType

Analysis

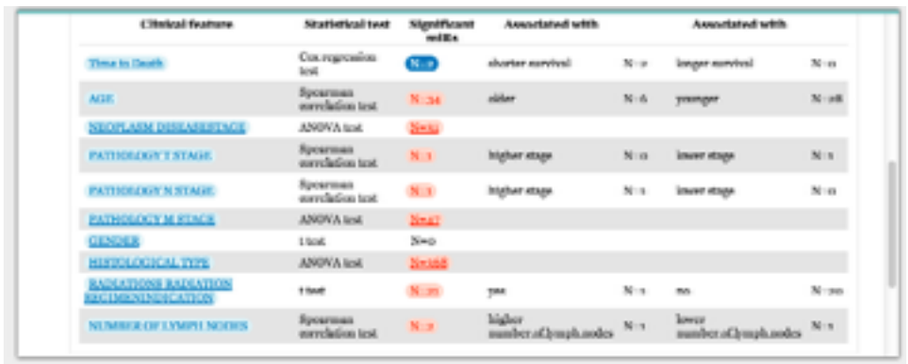
- Clustering Analyses
- Pathway Analyses
- Other Correlation Analyses

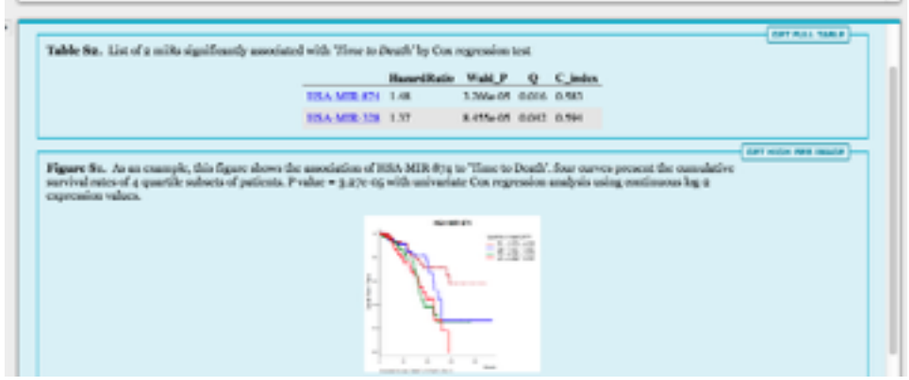
Correlation between miRseq expression and clinical features

Breast Invasive Carcinoma (Primary solid tumor)

16 April 2014 | analyses__2014_04_16 Maintainer Information Citation Information doi:10.7908/C1765CXV

Overview





H. Arachchi & Team

GDAC Firehose APIs

- 23 RESTful apis in 4 categories (more to come)
- Providing both bulk and fine-grained access
- Interactive docs : automatically updated as API evolves
- Automatically generated language bindings: Python, R, Matlab

Analyses : Fine grained retrieval of a

GET /Analyses/Mutation/MAF

GET /Analyses/Mutation/SMG

GET /Analyses/CopyNumber/Genes/All

GET /Analyses/CopyNumber/Genes/Focal

GET /Analyses/CopyNumber/Genes/Thresh

GET /Analyses/CopyNumber/Genes/Amplified

GET /Analyses/CopyNumber/Genes/Delete

GET /Analyses/Reports

GET /Analyses/Summary

Samples : Fine grained retrieval of sample-level data Show/Hide | List Operations

GET /Samples/mRNASeq

GET /Samples/miRSeq

GET /Samples/ClinicalTier1

Archives : Bulk retrieval of data or analysis pipeline results, as compressed archives

Show/Hide | List Operations | Expa

GET /Archives/StandardData Retrieve

Metadata : Retrieve disease, sample, and datatype descriptions, sample counts, and more

Show/Hide | List Operations | Expan

GET /Metadata/Counts Re

GET /Metadata/Cohorts Retrieve map of cohort abbreviation

GET /Metadata/Cohort/{cohort} Retrie

GET /Metadata/Platforms Retrieve map of platform code(s

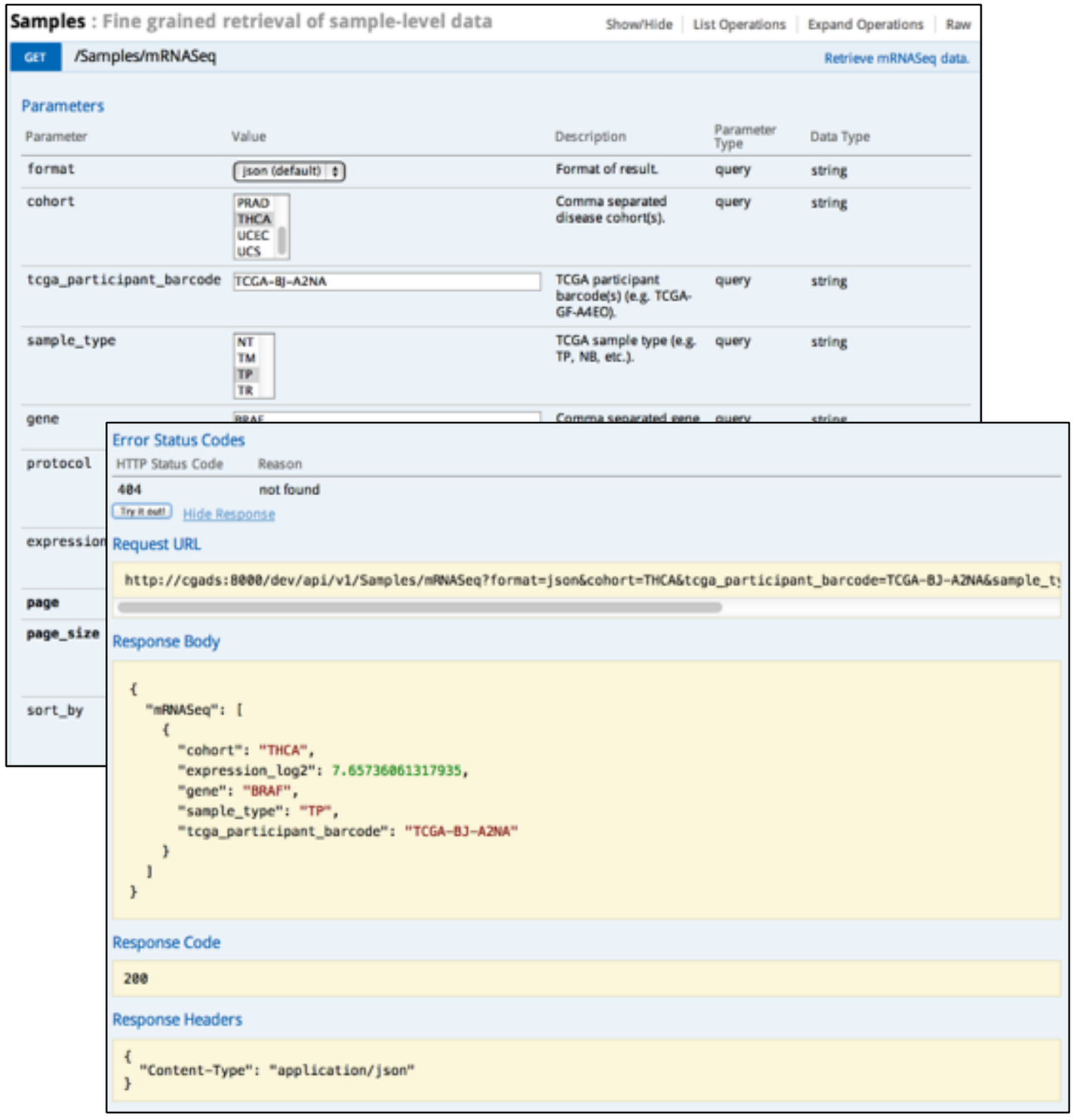
GET /Metadata/Centers Retrieve map of center name

Samples mRNASeq Example

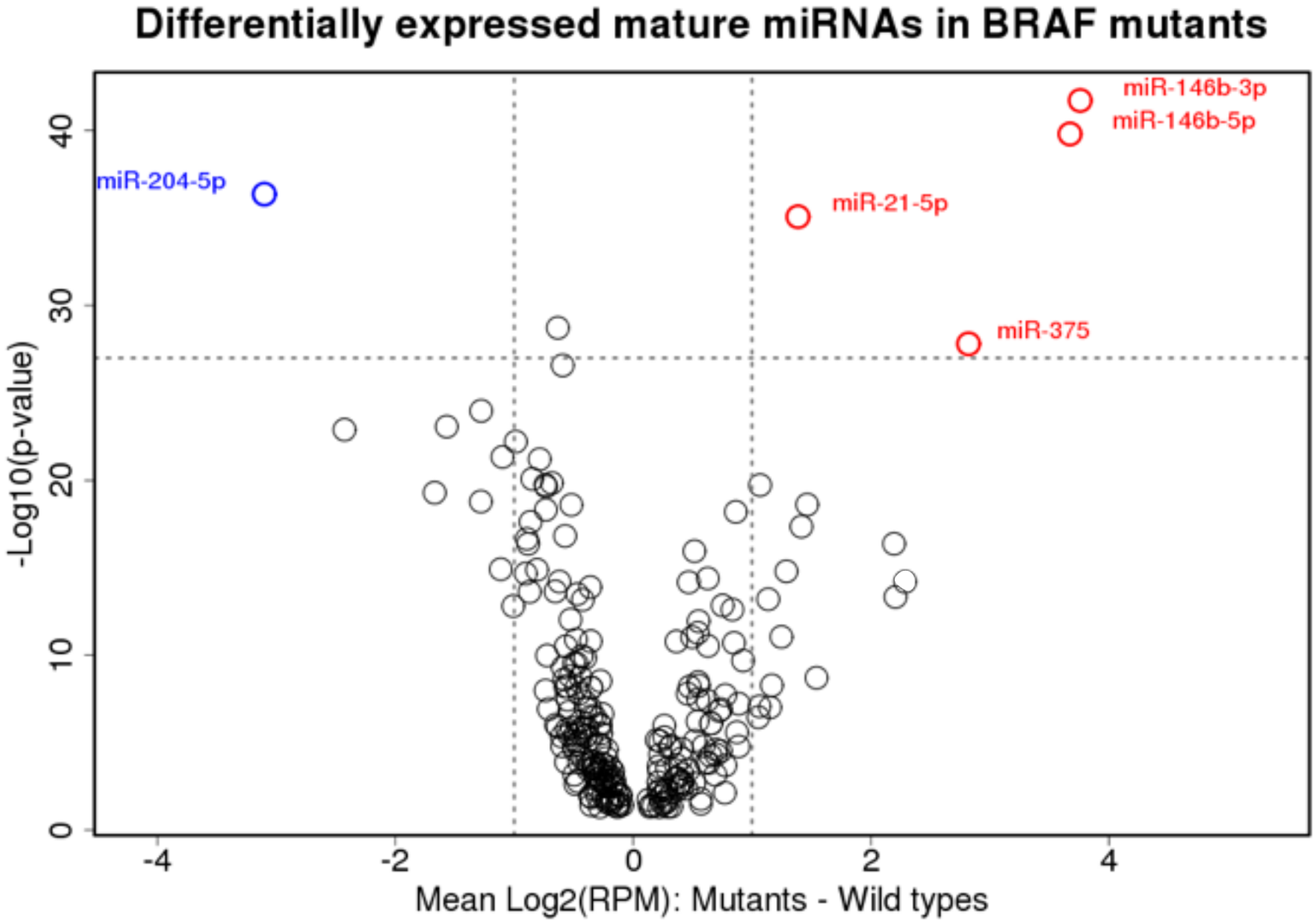
- Filters provide access to data of interest
  - Cohort (THCA, PRAD, etc.)
  - TCGA barcode (TCGA-BJ-A2NA)
  - Sample type (NT, NB, TP)
  - Gene (BRAF, NRAS)
  - Protocol (RSEM, RPKM)
  - Expression level threshold

- Obtain bulk or fine grained data
  - Bulk: obtain the entire tumor primary mRNASeq file for THCA
  - Fine grained: Obtain the RSEM estimated expression level of BRAF for THCA participant TCGA-BJ-A2NA (example displayed here)

- Retrieve results in JSON, TSV, or CSV



THCA Manuscript Reproducibility*



Generated in R + Python

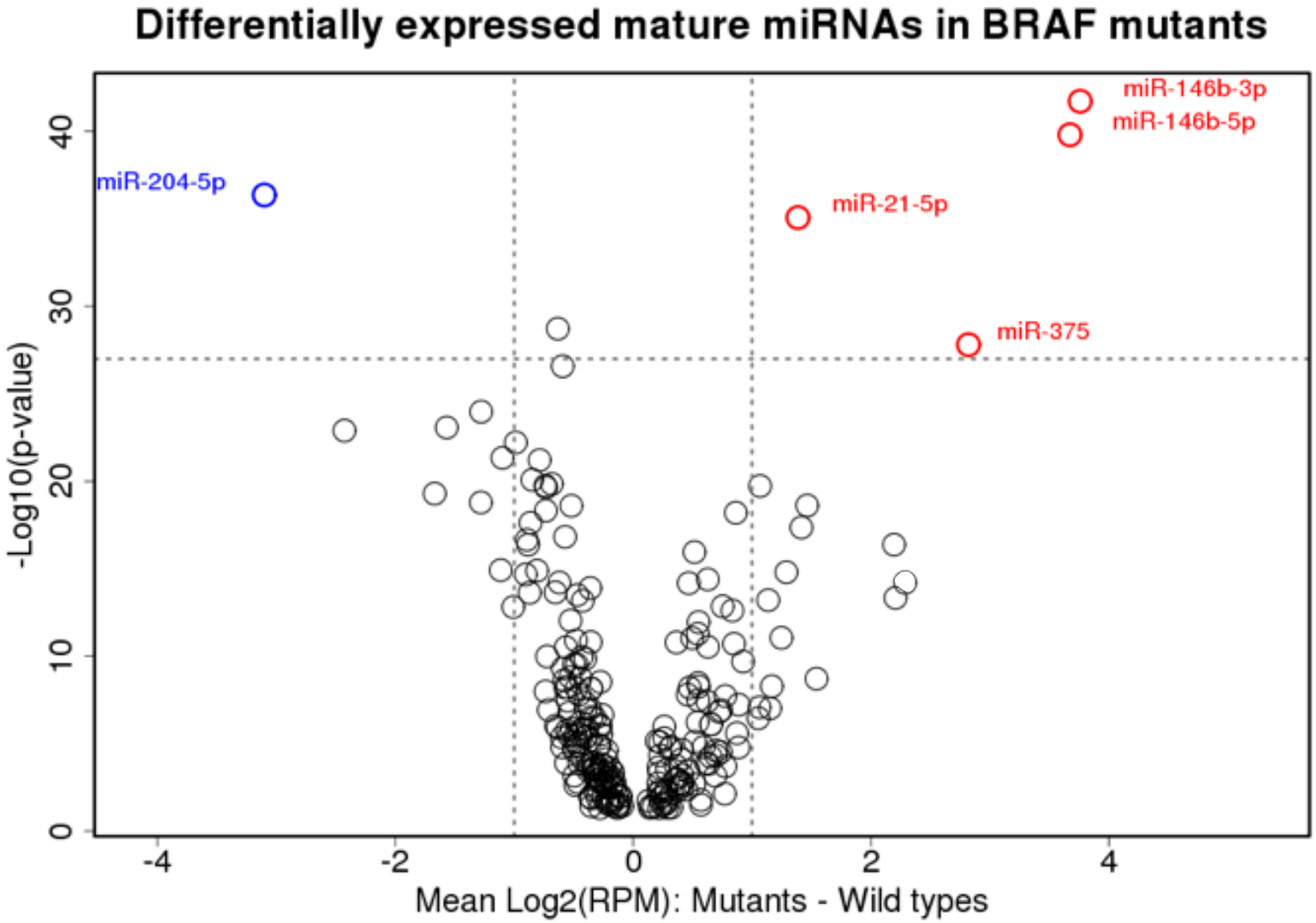
Data obtained
from Firehose API

MAF subset: retrieve only
needed columns

Retrieve subset of mature miRNAs

Jaegil Kim & David Heiman

THCA Manuscript Reproducibility*



Generated in R + Python

Data obtained
from Firehose API

MAF subset: retrieve only
needed columns

Retrieve subset of mature miRNAs

Figure from main text (submitted, see Giordano talk 9am tomorrow)

Jaegil Kim & David Heiman

Fin