



Broad GDAC

Lung Adenocarcinoma AWG Run 2013_02_07

Dan DiCara
Hailei Zhang
Michael Noble



GDAC Firehose Runs



- Firehose was conceived for the purpose of running a suite of analyses automatically on a monthly basis for public consumption and archival storage
- Evolved to include data standardization runs
 - Facilitates running automated analyses on a monthly basis with limited manual intervention
- Recent evolution to support AWGs directly
 - Running all Firehose/GDAC machinery on AWG data freezes and disease subtypes
 - Encapsulate results in browsable biologist friendly reports
 - Simplify data retrieval via `firehose_get`

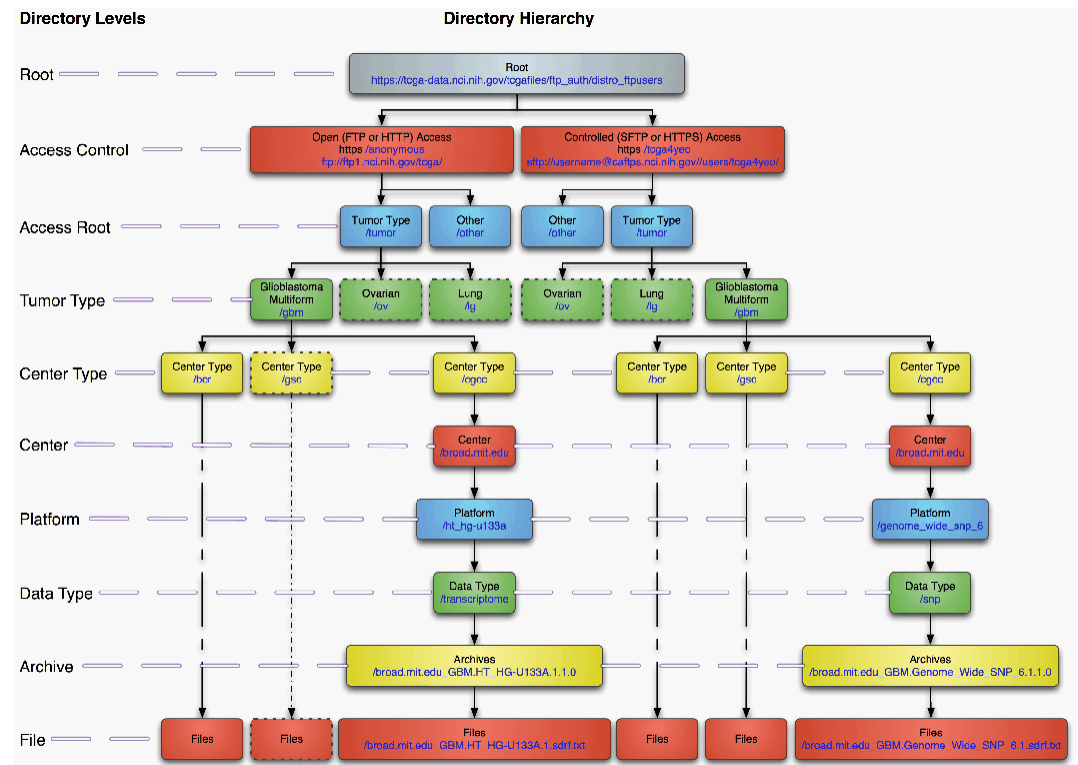
Outline



- GDAC Overview
 - Nightly mirroring and dicing
 - Workflows (Standard Data & Analyses)
 - Dashboards
 - Biologist Friendly Nozzle Reports
- Lung Adenocarcinoma AWG Run
 - Data freeze
 - Data retrieval via `firehose_get`
 - Analyses Results

Nightly Mirroring and Dicing

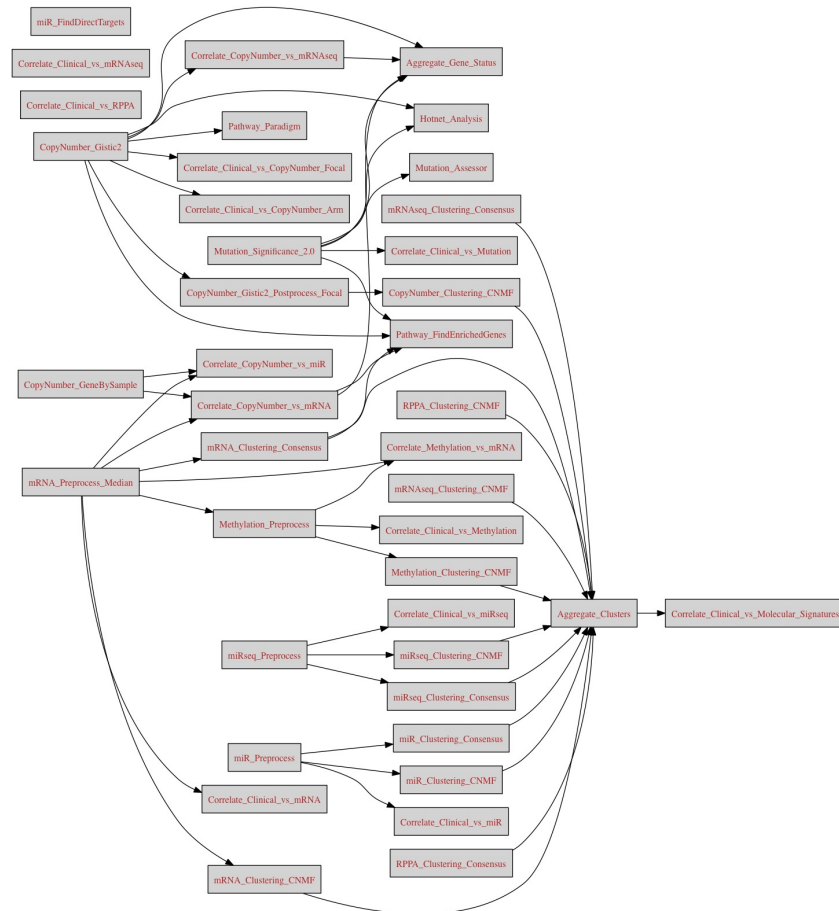
- Mirroring
 - Maintain an up-to-date snapshot of all available TCGA data at the DCC
- Dicing
 - Standardize data formatting to ensure it's amenable for automated analyses
 - Perform filtering
 - Redactions
 - Blacklisting
 - Replicate filtering



Workflows (Standard Data & Analyses)



- Standard data
 - 84 data merger pipelines
 - SDRFs generated to document data provenance
- Analyses
 - 41 analysis pipelines
 - Clustering
 - Correlation
 - Mutation
 - Copy number
 - Pathway
 - Biologist friendly result reports included



Dashboards



Home

Edit Share Add Tools

9 Added by Aaron Ball, last edited by Michael Noble on Dec 07, 2012 (view change) show comment

- [AWG Reps](#)
- [AWG Support](#)
- [Contact Us](#)
- [Dashboard-Analyses](#)
- [Dashboard-Stddata](#)
- [Data Usage Policy](#)
- [DCC Interactions](#)
- [Download](#)
- [FAQ](#)
- [Internal](#)
- [Nomenclature](#)
- [Nozzle](#)
- [Pipeline Docs](#)
- [Presentations](#)
- [ProcessFlow](#)
- [QualityControl](#)
- [Run Reports](#)
- [TAP](#)

[Email Archive](#)

[Tracking System](#)

Summary of 2013_01_16 stddata Run

DiseaseType	# Datasets	% Processed	Download
BLCA	78	100%	Open Protected
BRCA	117	100%	Open Protected
CESC	61	100%	Open Protected
COADREAD	93	100%	Open Protected
COAD	91	100%	Open Protected
DLBC	25	100%	Open Protected
GBM	119	100%	Open Protected
HNSC	83	100%	Open Protected
KICH	43	100%	Open Protected
KIRC	88	100%	Open Protected
KIRP	81	100%	Open Protected
LAML	28	100%	Open Protected
LGG	58	100%	Open Protected
LHC	66	100%	Open Protected
LUAD	104	100%	Open Protected
LUSC	112	100%	Open Protected
OV	153	100%	Open Protected
PAAD	48	100%	Open Protected
PRAD	62	100%	Open Protected
READ	93	100%	Open Protected
SARC	40	100%	Open Protected
SKCM	75	100%	Open Protected
STAD	58	100%	Open Protected
THCA	97	100%	Open Protected
UCEC	93	100%	Open Protected
PANCAN12	230	94%	Open Protected
PANCAN18	235	93%	Open Protected

Summary of 2013_01_16 stddata Run

DiseaseType	# Datasets	% Processed	Download
BLCA	78	100%	Open Protected
BRCA	117	100%	Open Protected
CESC	61	100%	Open Protected
COADREAD	93	100%	Open Protected
COAD	91	100%	Open Protected
DLBC	25	100%	Open Protected
GBM	119	100%	Open Protected
HNSC	83	100%	Open Protected
KICH	43	100%	Open Protected
KIRC	88	100%	Open Protected
KIRP	81	100%	Open Protected
LAML	28	100%	Open Protected
LGG	58	100%	Open Protected
LHC	66	100%	Open Protected
LUAD	104	100%	Open Protected
LUSC	112	100%	Open Protected
OV	153	100%	Open Protected
PAAD	48	100%	Open Protected
PRAD	62	100%	Open Protected
READ	93	100%	Open Protected
SARC	40	100%	Open Protected
SKCM	75	100%	Open Protected
STAD	58	100%	Open Protected
THCA	97	100%	Open Protected
UCEC	93	100%	Open Protected
PANCAN12	230	94%	Open Protected
PANCAN18	235	93%	Open Protected

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](#).

Bi-monthly Standard Data Results

2013_01_16 stddata Run

DiseaseType	# Datasets	% Processed	Download
BLCA	78	100%	Open Protected
BRCA	117	100%	Open Protected
CESC	61	100%	Open Protected
COADREAD	93	100%	Open Protected
COAD	91	100%	Open Protected
DLBC	25	100%	Open Protected
GBM	119	100%	Open Protected
HNSC	83	100%	Open Protected
KICH	43	100%	Open Protected
KIRC	88	100%	Open Protected
KIRP	81	100%	Open Protected
LAML	28	100%	Open Protected
LGG	58	100%	Open Protected
LHC	66	100%	Open Protected
LUAD	104	100%	Open Protected
LUSC	112	100%	Open Protected
OV	153	100%	Open Protected
PAAD	48	100%	Open Protected
PRAD	62	100%	Open Protected
READ	93	100%	Open Protected
SARC	40	100%	Open Protected
SKCM	75	100%	Open Protected
STAD	58	100%	Open Protected
THCA	97	100%	Open Protected
UCEC	93	100%	Open Protected
PANCAN12	230	94%	Open Protected
PANCAN18	235	93%	Open Protected

Monthly Analysis Results

2012_12_21 analyses Run

AnalysisReport	# Pipelines	% Successful	Download
BLCA	47	100%	Open Protected
BRCA	63	100%	Open Protected
CESC	44	100%	Open Protected
COADREAD	63	100%	Open Protected
COAD	63	100%	Open Protected
GBM	65	100%	Open Protected
HNSC	47	100%	Open Protected
KICH	24	100%	Open Protected
KIRC	63	100%	Open Protected
KIRP	60	100%	Open Protected
LGG	60	100%	Open Protected
LHC	16	100%	Open Protected
LUAD	63	100%	Open Protected
LUSC	63	100%	Open Protected
OV	69	100%	Open Protected
PAAD	21	100%	Open Protected
PRAD	44	100%	Open Protected
READ	63	100%	Open Protected
SARC	11	100%	Open Protected
SKCM	47	100%	Open Protected
STAD	42	100%	Open Protected
THCA	94	100%	Open Protected
UCEC	63	100%	Open Protected
LAML	12	92%	Open Protected
DLBC	6	86%	Open Protected
PANCAN12	6	38%	Open Protected
PANCAN18	4	31%	Open Protected

View [analysis reports](#) or click on [dashboards above](#) or download with [firehose get](#).

[November 2012 Firehose runs were not done, read here to understand why](#)

Dashboards cont.'d



Dashboard-Analyses

[Edit](#)
[Share](#)
[Add](#)
[Tools](#)

Added by [Michael Noble](#), last edited by [Michael Noble](#) on Jan 18, 2013 ([view change](#))

See [this presentation](#) on the role of Firehose within [The Cancer Genome Atlas](#), and note that downloading data from our site constitutes agreement to [this data usage policy](#).

2012_12_21 analyses Run

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

AnalysisReport	# Pipelines	% Successful	Download
BLCA	47	100%	Open Protected
BRCA	63	100%	Open Protected
CESC	44	100%	Open Protected
COADREAD	63	100%	Open Protected
COAD	63	100%	Open Protected
GBM	65	100%	Open Protected
HNSC	47	100%	Open Protected
KICH	24	100%	Open Protected
KIRC	63	100%	Open Protected
KIRP	60	100%	Open Protected
LGG	60	100%	Open Protected
LIHC	16	100%	Open Protected
LUAD	63	100%	Open Protected
LUSC	63	100%	Open Protected
OV	69	100%	Open Protected
PAAD	21	100%	Open Protected
PRAD	44	100%	Open Protected
READ	63	100%	Open Protected
SARC	11	100%	Open Protected
SKCM	47	100%	Open Protected
STAD	42	100%	Open Protected
THCA	94	100%	Open Protected
UCEC	63	100%	Open Protected
LAML	12	92%	Open Protected
DLBC	6	86%	Open Protected
PANCAN12	6	38%	Open Protected
PANCAN18	4	31%	Open Protected

[Analysis Reports](#)
[Download Data](#)

[View: Analysis reports](#)
[Release notes](#)
[FAQ](#)
[Download: firehose_get](#)

Dashboards cont.'d



**Summary of TCGA Tumor Data
Ingested into Broad GDAC Pipeline
2013_02_03 stddata Run**

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	153	117	152	105	153	0	122	0	135	54	28
BRCA	929	869	899	0	888	527	841	0	894	408	507
CESC	134	32	114	0	122	0	97	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	18	0	17	0	0	0	16	0	0
ESCA	20	0	0	0	20	0	0	0	0	0	0
GBM	598	565	563	0	405	542	161	491	0	214	291
HNSC	358	318	322	108	310	0	303	0	326	212	306
KICH	66	0	66	0	65	0	66	0	66	0	0
KIRC	502	502	493	0	500	72	480	0	481	454	293
KIRP	159	103	117	0	103	16	76	0	117	0	100
LAML	202	200	197	0	194	0	179	0	187	0	199
LGG	222	208	220	0	176	27	174	0	221	0	170
LIHC	99	62	97	0	98	0	34	0	96	0	0
LUAD	508	333	403	0	430	32	353	0	401	237	229
LUSC	399	327	358	0	359	154	258	0	349	195	178
OV	592	580	566	0	584	574	297	570	454	412	316
PAAD	57	0	57	0	49	0	31	0	34	0	34
PANCAN12	5345	4853	5021	423	4905	2179	3415	1061	4264	2785	2819
PRAD	190	148	177	0	172	0	140	0	177	0	83
READ	169	168	162	35	162	69	72	0	143	130	69
SARC	52	0	29	0	29	0	0	0	29	0	0
SKCM	288	184	273	119	288	0	265	0	272	164	253
STAD	308	162	237	0	257	0	43	0	237	0	116
THCA	500	287	430	94	435	0	379	0	411	224	323
UCEC	512	451	493	106	500	54	370	0	487	200	248
Totals	7468	6039	6856	636	6736	2222	4933	1061	6062	3173	3934

Dashboards cont.'d



UP < > EXPAND ALL COLLAPSE ALL SET AUTO WIDTH PRINT

stddata__2013_02_03 Samples Summary Report

- Overview
- Introduction
- + Summary
- Results
- + Redactions
- + Replicate Filtered Samples
- + Blacklisted Samples
- Sample Heatmaps

BLCA

Figure 1. This figure depicts the distribution of available data on a per participant basis.

GET HIGH-RES IMAGE

Dashboards cont.'d



Dashboard-Analyses

[Edit](#) [Share](#) [Add](#) [Tools](#)

Added by [Michael Noble](#), last edited by [Michael Noble](#) on Jan 18, 2013 ([view change](#))

See [this presentation](#) on the role of Firehose within [The Cancer Genome Atlas](#), and note that downloading data from our site constitutes agreement to [this data usage policy](#).

2012_12_21 analyses Run

Tables of Ingested Data: [HTML](#) [PNG](#) [TSV](#) | Samples Summary: [Report](#)

Sample Counts Table (points to the table)

Samples Summary Report (points to the Report link)

Analysis Reports (points to the LUAD row)

Download Data (points to the Protected link in the LUAD row)

Prior Analysis Runs

- [Oct 24, 2012](#)
- [Sept 13, 2012](#)
- [Aug 25, 2012](#)
- [July 25, 2012](#)
- [June 23, 2012](#)
- [May 25, 2012](#)
- [April 25, 2012](#)
- [March 21, 2012](#)
- [Feb 17, 2012](#)
- [Jan 24, 2012](#)
- [Dec 30, 2011](#)
- [Nov 28, 2011](#)
- [Oct 26, 2011](#)
- [Sept 21, 2011](#)
- [July 28, 2011](#)
- [May 25, 2011](#)
- [April 21, 2011](#)
- [March 27, 2011](#)
- [Feb 17, 2011](#)

AnalysisReport	# Pipelines	% Successful	Download
BLCA	47	100%	Open Protected
BRCA	63	100%	Open Protected
CESC	44	100%	Open Protected
COADREAD	63	100%	Open Protected
COAD	63	100%	Open Protected
GBM	65	100%	Open Protected
HNSC	47	100%	Open Protected
KICH	24	100%	Open Protected
KIRC	63	100%	Open Protected
KIRP	60	100%	Open Protected
LGG	60	100%	Open Protected
LIHC	16	100%	Open Protected
LUAD	63	100%	Open Protected
LUSC	63	100%	Open Protected
OV	69	100%	Open Protected
PAAD	21	100%	Open Protected
PRAD	44	100%	Open Protected
READ	63	100%	Open Protected
SARC	11	100%	Open Protected
SKCM	47	100%	Open Protected
STAD	42	100%	Open Protected
THCA	94	100%	Open Protected
UCEC	63	100%	Open Protected
LAML	12	92%	Open Protected
DLBC	6	86%	Open Protected
PANCAN12	6	38%	Open Protected
PANCAN18	4	31%	Open Protected

View: [Analysis reports](#) [Release notes](#) [FAQ](#) | Download: [firehose_get](#)

Biologist Friendly Nozzle Reports



Analysis Overview for Lung Adenocarcinoma

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

Overview

Introduction

This is the analysis overview for Firehose run "21 December 2012".

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 356 tumor samples used in this analysis: 27 significant arm-level results, 30 significant focal amplifications, and 45 significant focal deletions were found.

Mutation Analysis (MutSig v2.0)

[View Report](#) |

Mutation Analysis (MutSig vS2N)

[View Report](#) |

Clustering Analyses

Clustering of copy number data: consensus NMF

[View Report](#) | The most robust consensus NMF clustering of 356 samples using the 75 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 3800 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 304 samples and 3800 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 RPPA data: consensus NMF

[View Report](#) | The most robust consensus NMF clustering of 237 samples using the 150 most variable proteins 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 RPPA data: consensus hierarchical

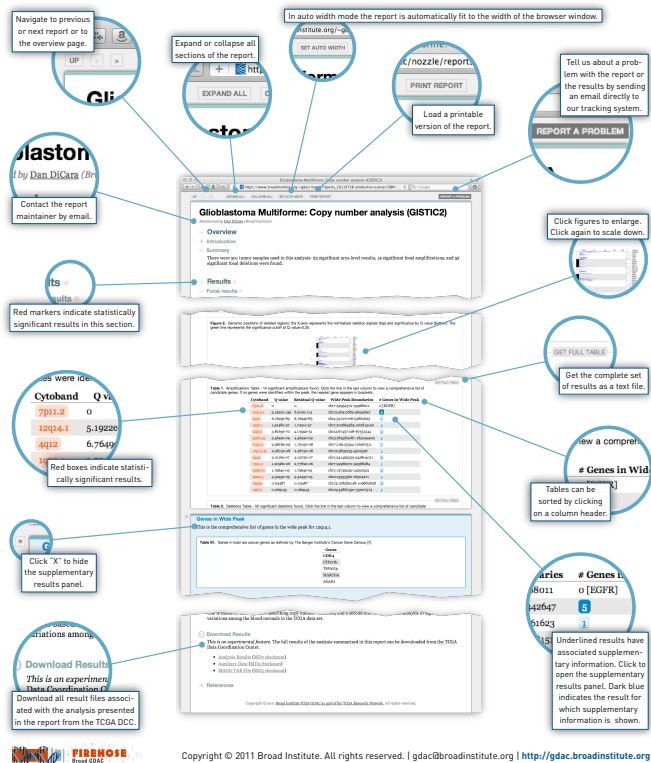
[View Report](#) | The 150 most variable proteins were selected. Consensus average linkage hierarchical clustering of 237 samples and 150 proteins identified 4 subtypes with the stability of the clustering increasing for $k = 2$ to $k = 8$ and the average silhouette width calculation for selecting the robust clusters.

Clustering of mRNA expression: consensus NMF

Biologist Friendly Nozzle Reports cont.'d

Firehose Reports | At-a-Glance

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



Navigation: UP, DOWN, EXPAND ALL, COLLAPSE ALL, SET AUTO WIDTH, PRINT

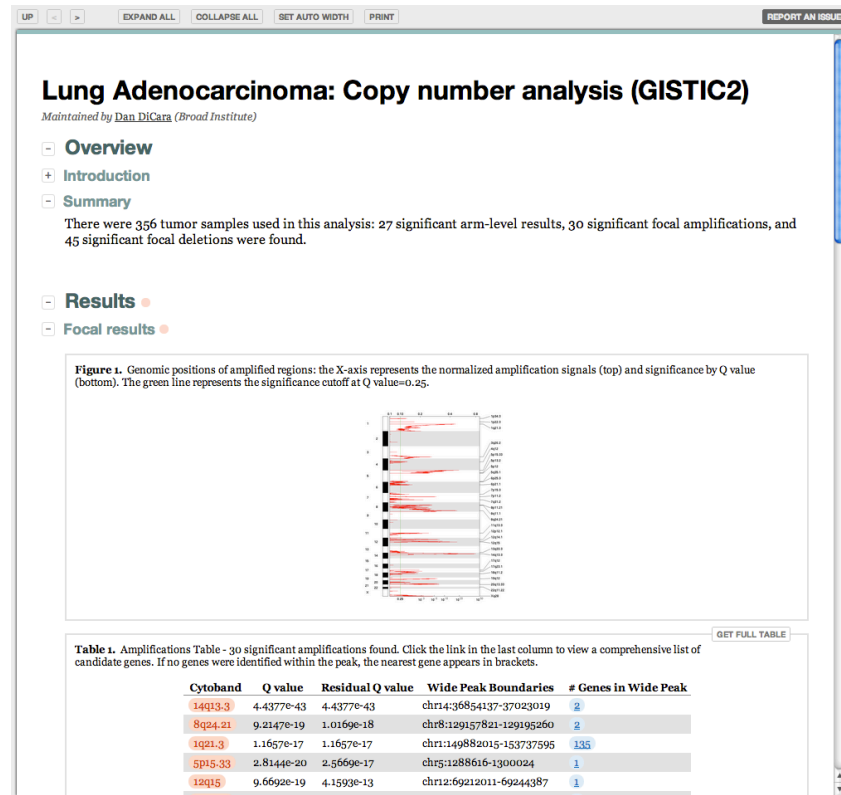
Interactions: Click figures to enlarge, Click again to scale down. GET FULL TABLE. Get the complete set of results as a text file. View a comprehensive table of results. Tables can be sorted by clicking on a column header. Underlined results have associated supplementary information. Click to open the supplementary results panel. Dark blue indicates the result for which supplementary information is shown.

Reporting: Tell us about a problem with the report or the results by sending an email directly to our tracking system. Lead a printable version of the report.

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

Download: Download all result files associated with the analysis presented in the report from the TCGA DCC.

Copyright © 2011 Broad Institute. All rights reserved. | gdc@broadinstitute.org | http://gdc.broadinstitute.org



Lung Adenocarcinoma: Copy number analysis (GISTIC2)

Maintained by [Dan DiCara](#) (Broad Institute)

- Overview
- Introduction
- Summary

There were 356 tumor samples used in this analysis: 27 significant arm-level results, 30 significant focal amplifications, and 45 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.

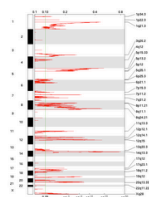


Table 1. Amplifications Table - 30 significant amplifications found. Click the link in the last column to view a comprehensive list of candidate genes. If no genes were identified within the peak, the nearest gene appears in brackets.

Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
14q13.3	4.4377e-43	4.4377e-43	chr14:36854137-37023019	2
8q24.21	9.2147e-19	1.0169e-18	chr8:129157821-129195260	2
1q21.3	1.1657e-17	1.1657e-17	chr1:49882015-153737595	135
5p15.33	2.8144e-20	2.5659e-17	chr5:1288616-1300024	1
12q15	9.6692e-19	4.1593e-13	chr12:69212011-69244387	1

GET FULL TABLE

Data Freeze

- Standard data and analyses runs performed on a frozen subset of samples
- Subtype analyses possible (i.e. run the entire GDAC pipeline on each subtype)
 - Molecular Smoker
 - Molecular Non-smoker
 - Oncogene Positive
 - ...

Analysis Overview for 07 February 2013

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

Unique Tumor Sample Counts

Tumor	BCR	Clinical	CN	Methylation	mRNA	mRNASeq	miRSeq	RPPA	MAF
luad	230	229	230	229	23	230	230	181	129

Download run results with [firehose_get](#)

Simplest download command: `firehose_get awg_luad 2013_02_07`

For More Help: `firehose_get --help`

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.

A total of 199 reports are available for analysis run "07 February 2013".

Results

Cancer Types

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

Cancer Type	Cohort	Reports	HTML	ZIP
Lung Adenocarcinoma	LUAD-BRONCHIOID	26	Browse	Download
Lung Adenocarcinoma	LUAD-MAGNOID	30	Browse	Download
Lung Adenocarcinoma	LUAD-MOLECULAR_NONSMOKER	30	Browse	Download
Lung Adenocarcinoma	LUAD-MOLECULAR_SMOKER	31	Browse	Download
Lung Adenocarcinoma	LUAD-ONCOGENE_NEGATIVE	26	Browse	Download
Lung Adenocarcinoma	LUAD-ONCOGENE_POSITIVE	30	Browse	Download
Lung Adenocarcinoma	LUAD-SQUAMOID	26	Browse	Download

An archive containing all reports for the 07 February 2013 run is available for [download](#). Links to archives with reports for individual cancer types are included in Table 1.

http://gdac.broadinstitute.org/runs/awg_luad_2013_02_07/reports

Data Retrieval via firehose_get



- Central location for collaborators to retrieve data and analyses for an AWG freeze
 - Easy data retrieval:
firehose_get awg_luad
2013_02_07
 - Ensures consistency across all centers analyzing TCGA data for an AWG
 - Easier than passing around a freeze list and having each analyst curate his/her own frozen data

```
3:40pm dicara@cga03 ~ $ firehose_get -help
firehose_get : retrieve open-access results of Broad Institute TCGA GDAC runs
Version: 0.3.10 (Author: Michael S. Noble)

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

Two arguments are required; the first must be one of

    analyses  awg_lgg  awg_luad  awg_pancan8
    awg_skcml awg_thca  stddata

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

    BLCA  BRCA  CESC  COAD  COADREAD  DLBC  ESCA  GBM  HNSC  KICH
    KIRC  KIRP  LAML  LGG  LIHC  LUAD  LUSC  OV  PAAD  PANCAN
    PANCAN8  PANCAN12  PRAD  READ  SARC  SKCM  STAD  THCA  UCEC

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

Flags:

-b | -batch          do not prompt: assume YES answer to all queries
-c | -cohorts        list available disease cohorts
-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
-p | -platforms      list data platforms available in Firehose runs
                        (not implemented yet)
-r | -runs           list available Firehose runs
-t | -tasks <list>  further prune the set of archives retrieved, by
                        INCLUDING only the tasks (pipelines) whose
                        names match the given space-delimited list of
                        patterns; matching is performed with glob-style
                        wildcards; if a tilde ~ is prepended to a task
                        name then matching tasks will be EXCLUDED; when
                        no pattern list is given firehose_get will display
                        all tasks in the selected run

NOTE: not all tasks will execute for all disease
cohorts; what tasks are run depends upon the
data available for that disease cohort

-v          display the version of firehose_get
-x          debugging: turn on bash set -x (warning: very verbose)

Broad GDAC website: http://gdac.broadinstitute.org
Broad GDAC email  : gdac@broadinstitute.org

3:41pm dicara@cga03 ~ $ |
```

Analyses Results Outline



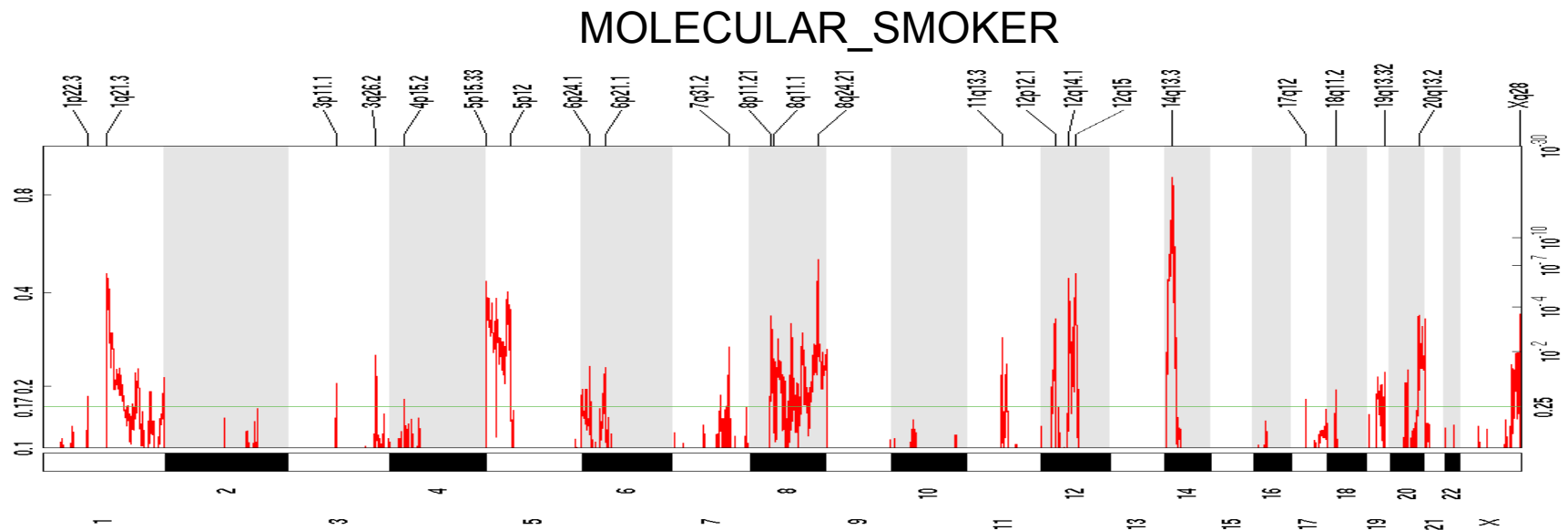
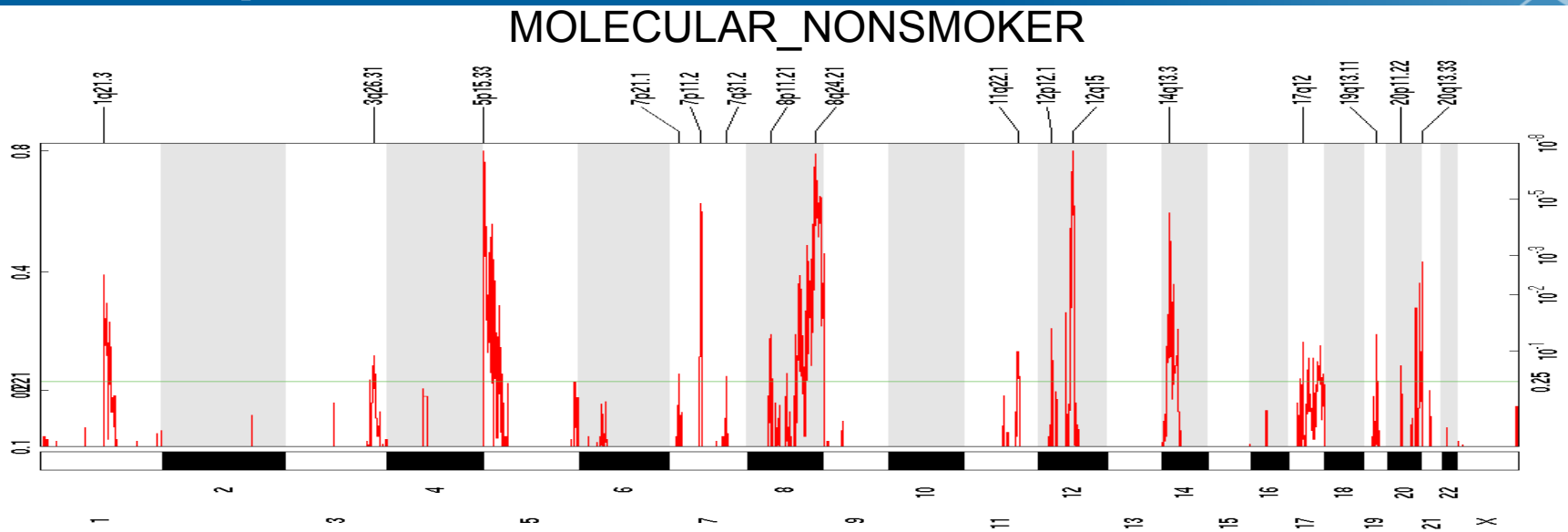
Subtypes	No. patients
Molecular_Nonsmoker	81
Molecular_Smoker	149
Oncogene_Negative	96
Oncogene_Positive	134
Bronchioid	89
Magnoid	63
Squamoid	78

The analysis include copy number GISTIC, Mutsig (v2.0 & s2N), clustering for expression platforms, Integrated with clinical info, Pathway analysis– Hotnet and

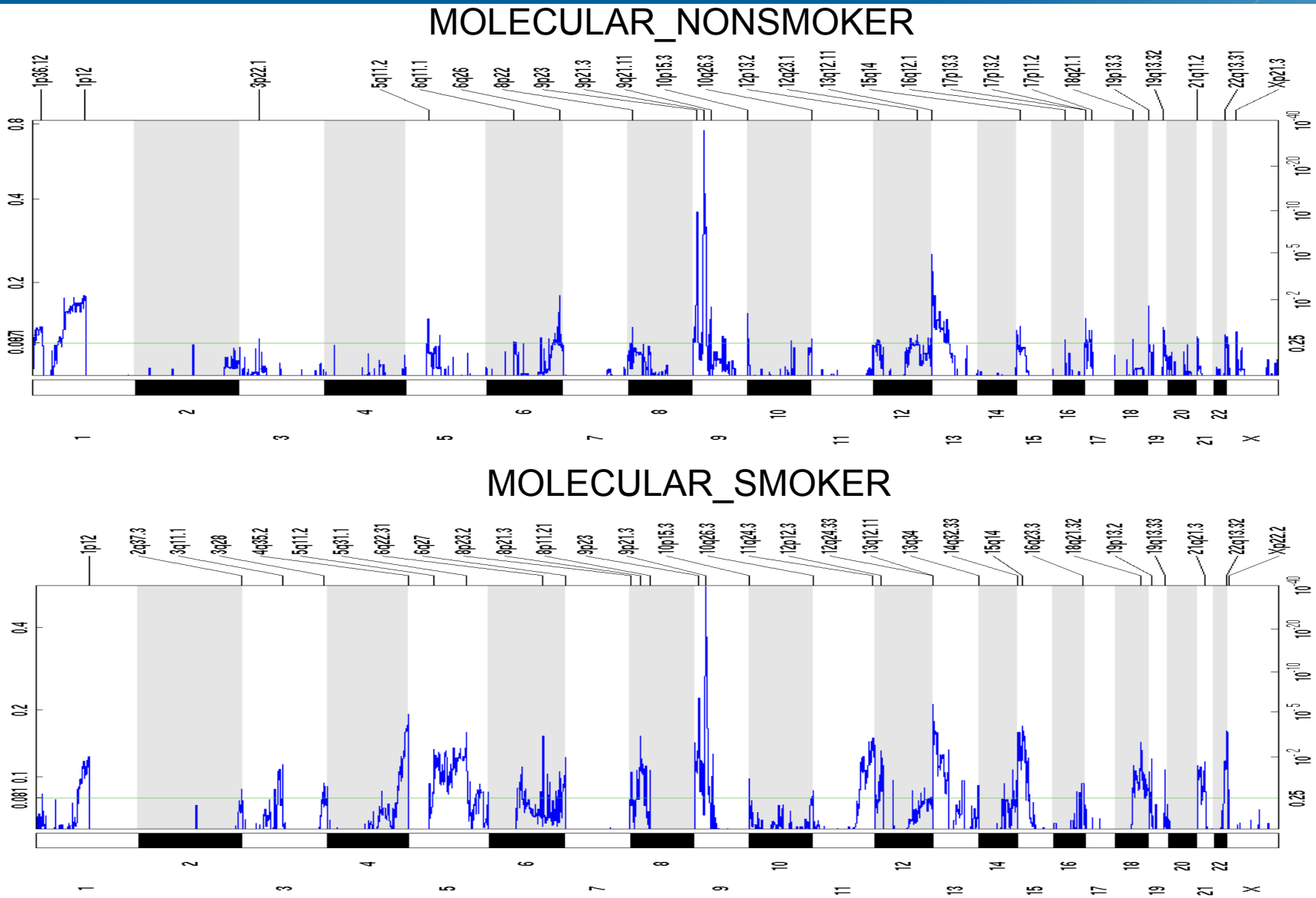
IGV view of copy number -- Nonsmokers and smokers



Copy number analysis (GISTIC 2) – Focal Amplifications



Copy number analysis (GISTIC 2) – Focal Deletions



Nonsmoker--MutSig 2.0 and S2N



Mutsig S2N			Mutsig v2.0		
gene	p	q	gene	p	q
TP53	0	0	TP53	1.1e-14	2e-10
SETD2	5.9e-32	5.6e-28	EGFR	1e-11	9.1e-08
EGFR	1.7e-27	1.1e-23	CDKN2A	5.4e-08	0.00033
BRAF	1.8e-22	8.4e-19	SMAD4	1.6e-07	0.00072
EIF5B	3.3e-19	1.2e-15	KRAS	3.7e-07	0.0013
CDKN2A	5.9e-17	1.8e-13	KEAP1	5.9e-07	0.0018
SMAD4	2.9e-10	7.8e-07	CSMD3	1e-06	0.0027
KEAP1	6.3e-10	1.5e-06	BRAF	0.000018	0.041
GLG1	1.3e-08	0.000028	SPTA1	0.000034	0.063
KRAS	3.5e-06	0.0065	STK11	0.000035	0.063
LONP1	0.000016	0.028	OR4A5	0.000055	0.09

* No mutation genes are correlated with clinical info.

Smoker--MutSig 2.0 and S2N



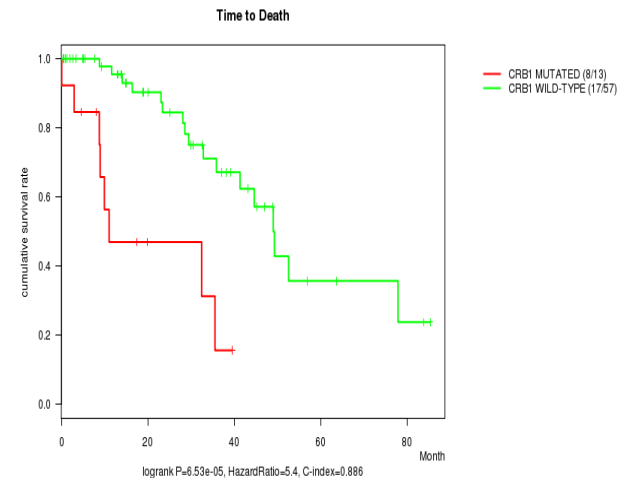
Mutsig S2N

gene	p	q
TP53	2.9e-82	5.6e-78
KRAS	1.1e-47	1.1e-43
STK11	2.9e-35	1.8e-31
RBM10	3.2e-30	1.5e-26
FAM75C1	4e-21	1.5e-17
ZNF770	6.4e-12	2e-08
AGAP6	1.9e-08	0.000051
ZNF679	2.7e-08	0.000064
FSCB	2.3e-07	0.00047
HRNR	3.5e-06	0.0067
KEAP1	8.9e-06	0.015

Mutsig v2.0 – top 10

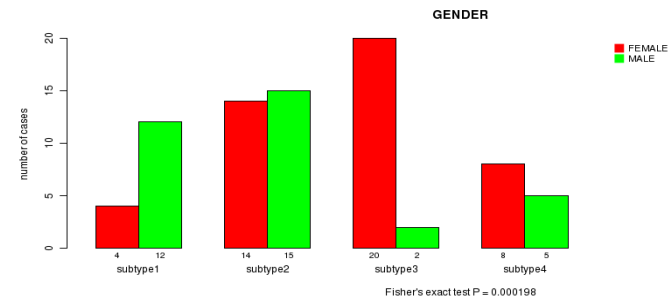
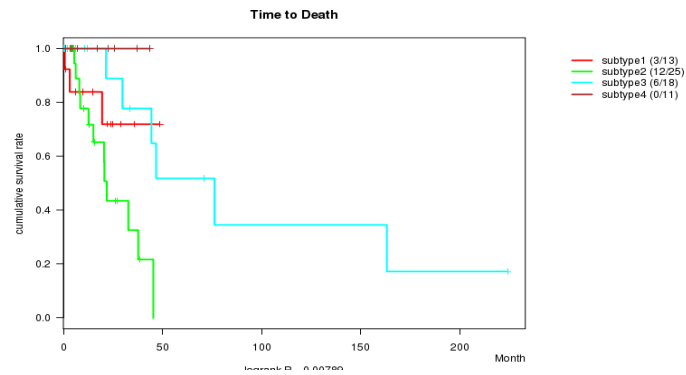
gene	p	q
KRAS	2.78E-15	2.71E-11
TP53	4.55E-15	2.71E-11
KEAP1	5.77E-15	2.71E-11
STK11	6.00E-15	2.71E-11
REG3A	9.02E-10	3.26E-06
NAV3	8.56E-09	2.58E-05
SNTG1	2.06E-08	4.84E-05
SLITRK2	2.14E-08	4.84E-05
OR2T33	2.44E-08	4.90E-05
RIMS2	2.83E-08	5.12E-05
CD5L	4.31E-08	7.08E-05

CRB1

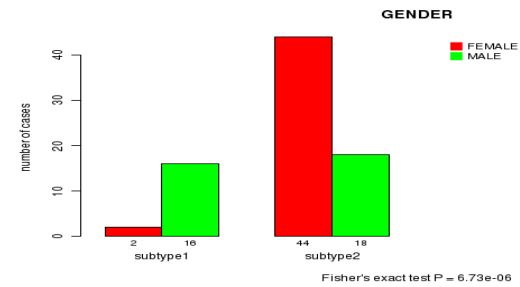
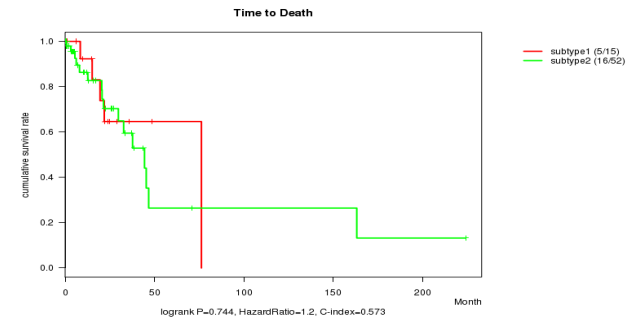


Nonsmoker--mRNAseq gene expression clustering

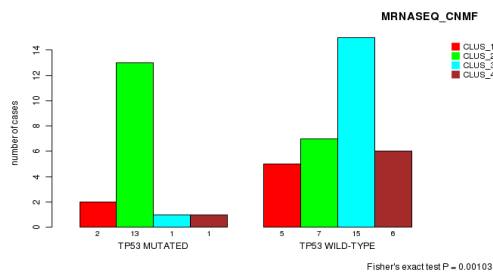
cNMF



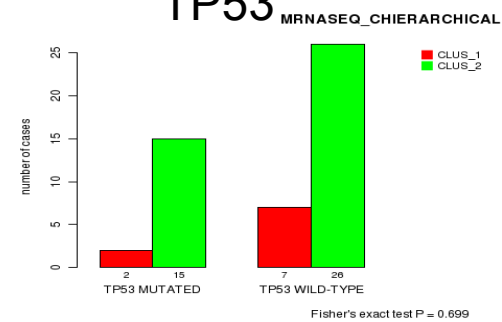
cHierarchical



TP53



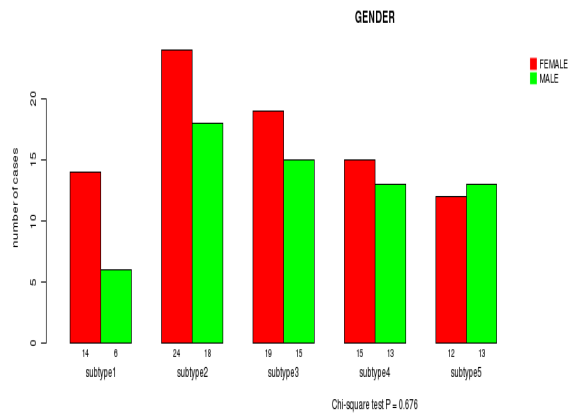
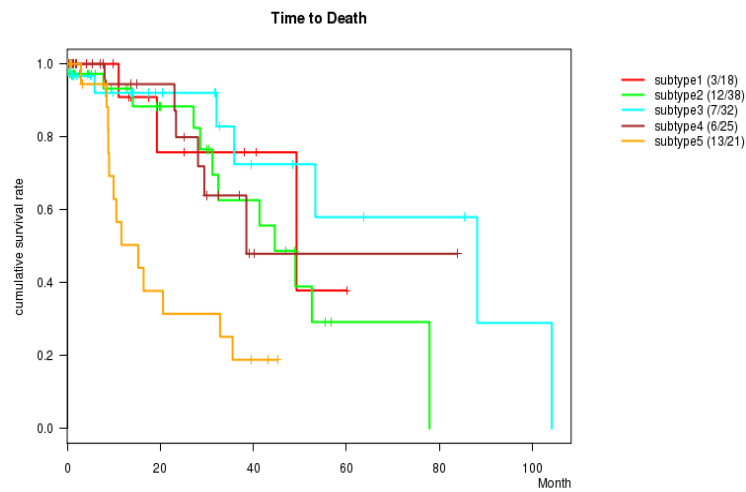
TP53



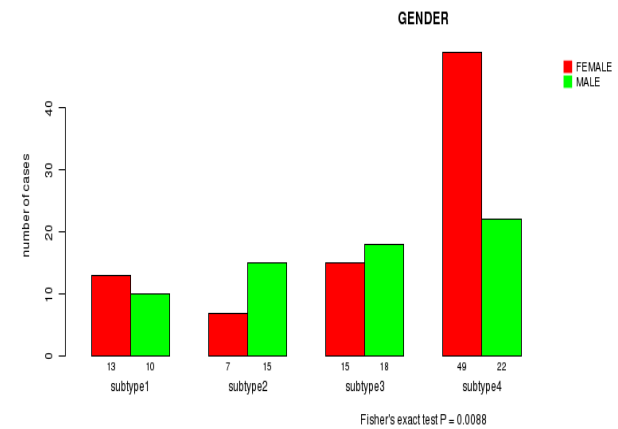
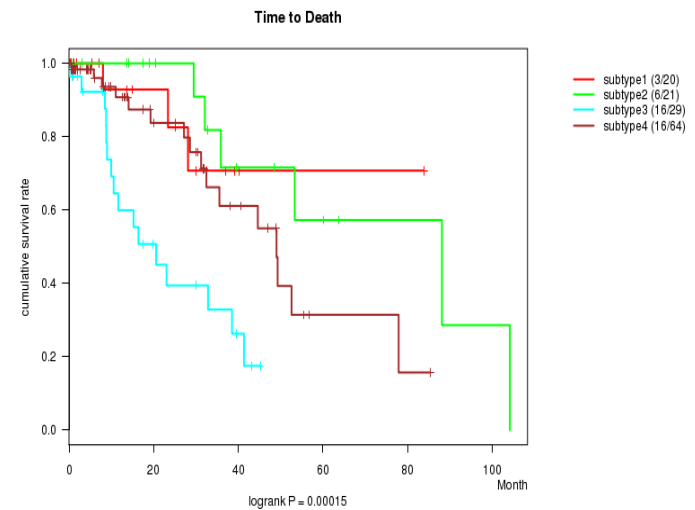
Smoker --- mRNAseq gene expression clustering



cNMF



cHierarchical



PARADIGM pathway analysis of mRNASeq expression and copy number data



MOLECULAR_NONSMOKER

Pathway.Name	Avg.Num.Perturbations
IL4-mediated signaling events	11
FOXA2 and FOXA3 transcription factor networks	11
Signaling events mediated by the Hedgehog family	9
Signaling mediated by p38-alpha and p38-beta	9
Syndecan-4-mediated signaling events	7
Osteopontin-mediated events	6
HIF-1-alpha transcription factor network	6
Endothelins	6
EGFR-dependent Endothelin signaling events	6
Nephrin/Neph1 signaling in the kidney podocyte	5

MOLECULAR_SMOKER

Pathway.Name	Avg.Num.Perturbations
Signaling events mediated by the Hedgehog family	20
IL4-mediated signaling events	20
Signaling mediated by p38-alpha and p38-beta	18
Endothelins	16
HIF-1-alpha transcription factor network	14
FOXA2 and FOXA3 transcription factor networks	14
Syndecan-1-mediated signaling events	13
Wnt signaling	13
IL23-mediated signaling events	13
Visual signal transduction: Cones	12