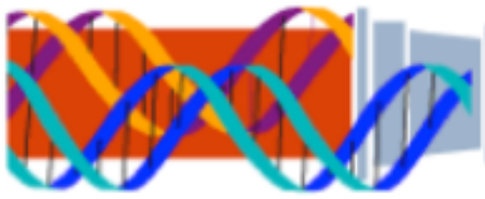


FIREBROWSE:

Mining the Firehose of TCGA Genomic Data

Michael S. Noble
Associate Director, Cancer Genome Analysis
Broad Institute of MIT and Harvard

Boston, Massachusetts, USA
April 6, 2016

BROAD INSTITUTE

THE CANCER GENOME ATLAS

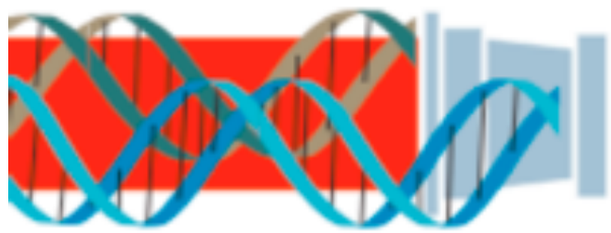
MASSACHUSETTS GENERAL HOSPITAL CANCER CENTER

A brief history is helpful to put FireBrowse in context

Helped catalyze a new era:

Collaborative Science @ Extreme Scale

We are privileged to bear witness to the transformation of an entire field, biomedical research: from largely wet & qualitative ⟶ highly digital & quantitative

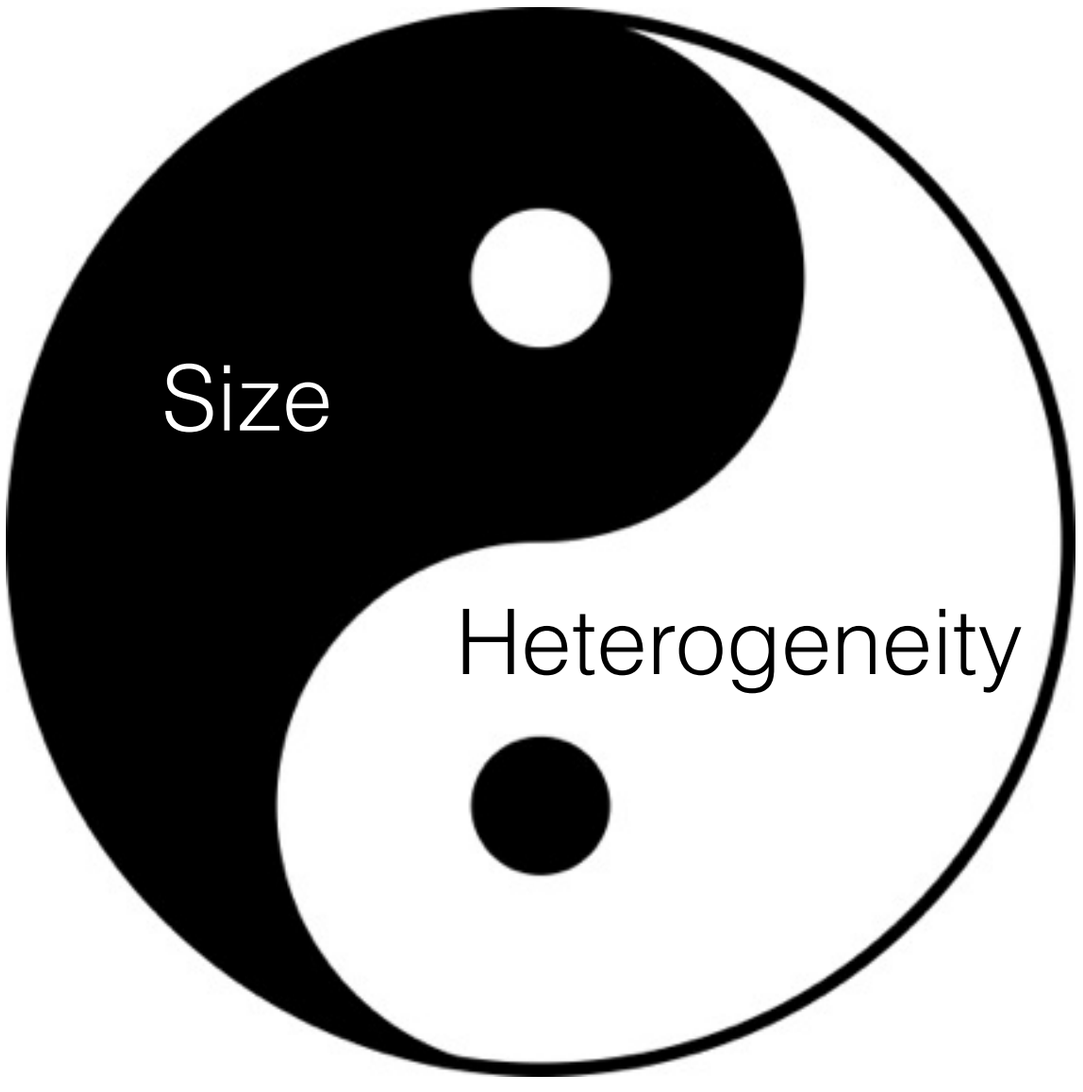


FIREHOSE
Broad GDAC

Born of the desire to systematize analyses from TCGA pilot and scale their execution to the dozens of remaining diseases to be studied.

Now sits atop >50 TB of analysis-ready TCGA data, and reliably executes thousands of pipelines per month.

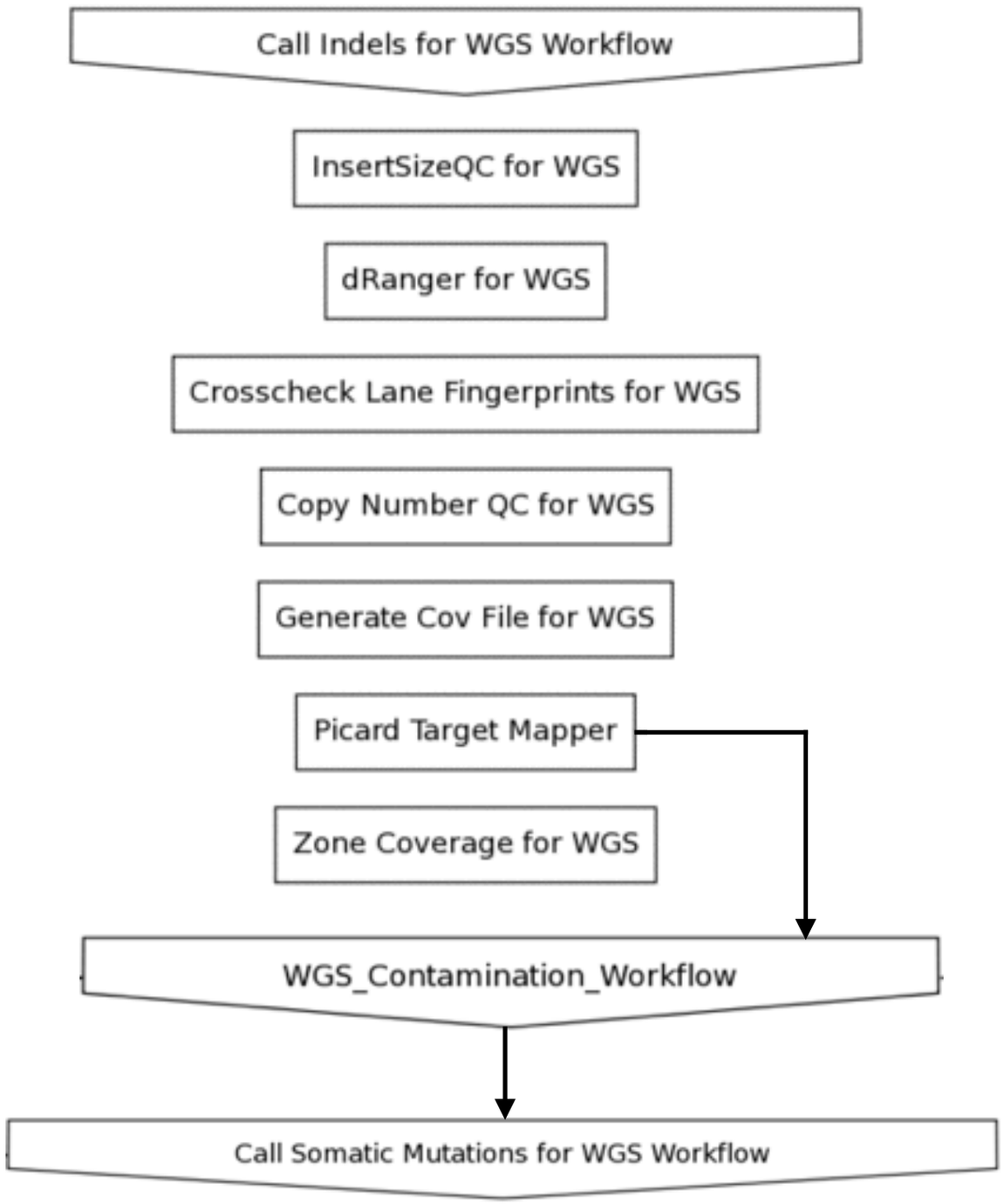
You may think 50 TB is ~small these days



But size and heterogeneity are dualities of the same problem: complexity

… and scaling for size is arguably easier

Exhibit A: whole genome mutation calling



Input: huge BAM files

Essentially linear

Most of ~1.5 PB of TCGA data are BAMs:

Huge, but simple-ish knowledge content:

3x10^9 A,T,C,G

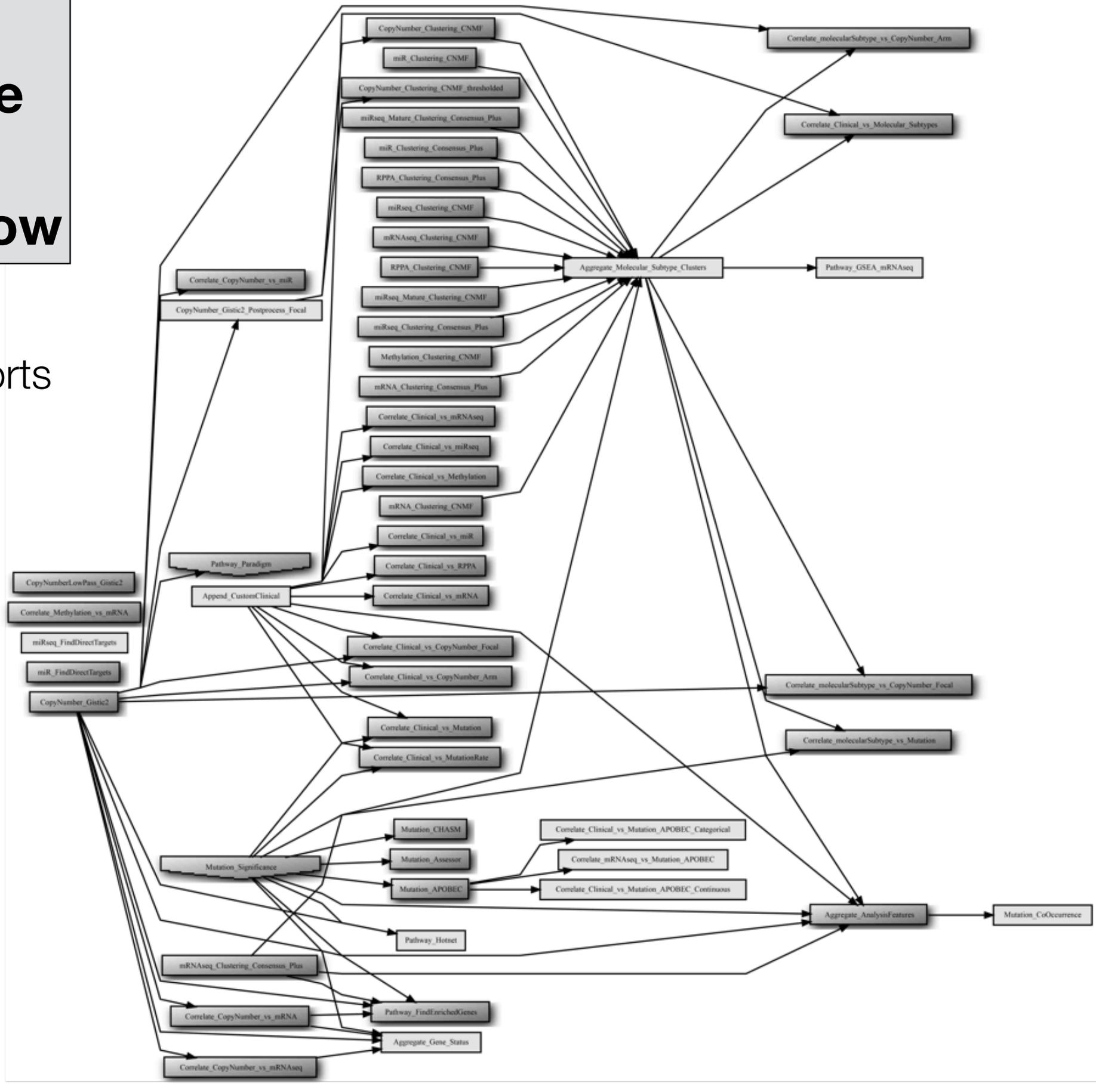
Exhibit B: GDAC Firehose Integrated Analysis Workflow

Run on all TCGA cohorts
>100 tasks per

Wiring much more complex

Inputs much smaller

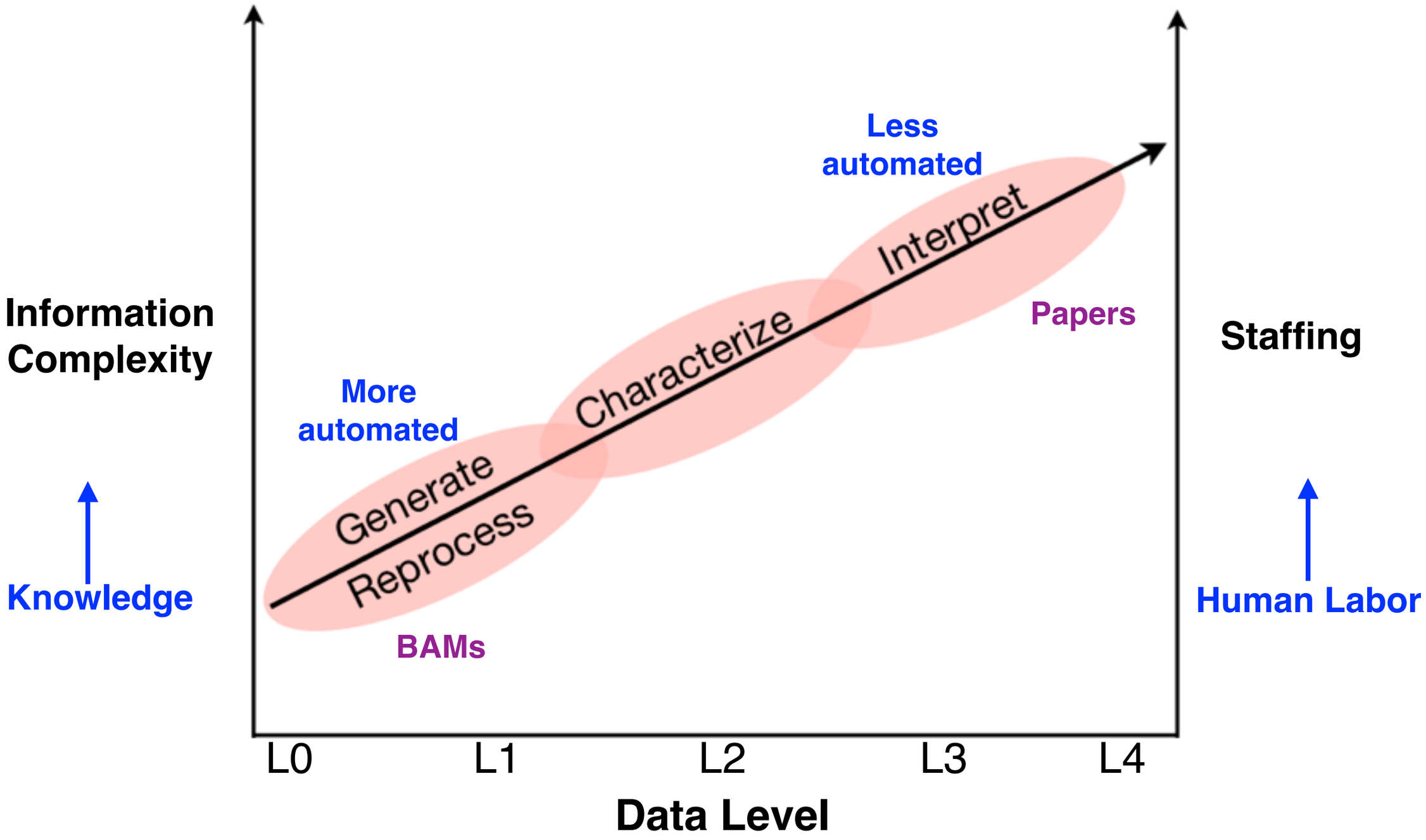
But far more heterogeneous:
10 input data types



Which do you think is harder to keep running?

Wrestling with enormous complexity

- pipelines are compression algorithms: derived files get smaller
- but greatly increase in number & semantic diversity:
  copy number, mutation, expression, protein, methylation, etc
- knowledge encoded per byte goes up
- as does need for direct human involvement



The “big” in *Big Data* gets lots of buzz, but *size alone* is often not the biggest problem

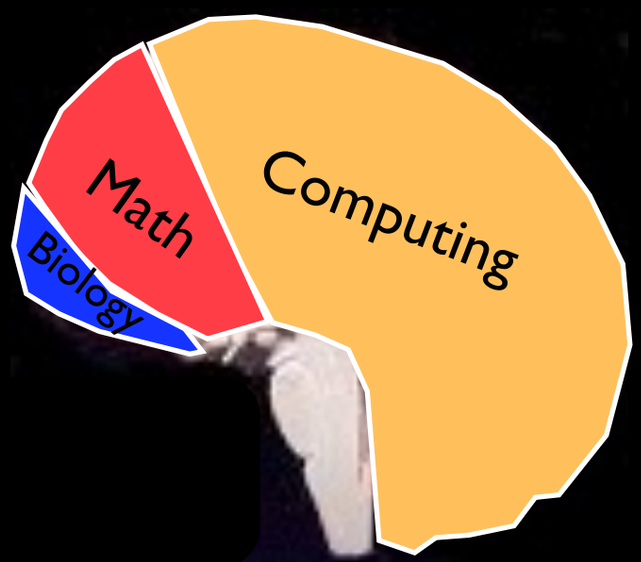
Complexity can be a silent, incremental and harder to please companion

Discovery is emergent:
data & results must be integrated in increasingly complex ways,not analyzed in isolation, for insight to happen

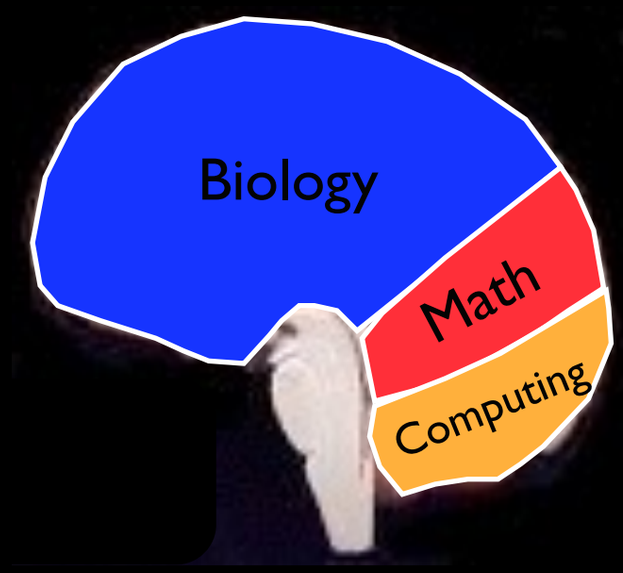
Poorly managed, complexity stifles science

This is Your
Researcher
Brain

Poorly managed, complexity stifles science



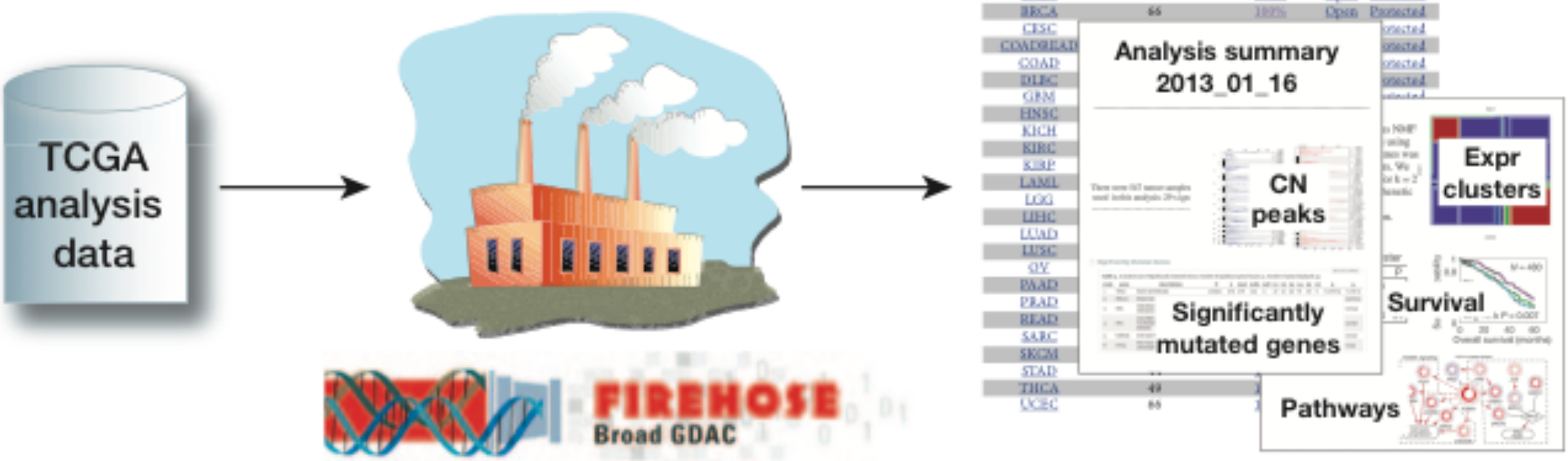
When Coding
Or Data
Exploration
Is Hard or
Untrusted



When
Easier

Acute Need for Automation, Systematic Rigor, and Transparency

Data Factory



gdac.broadinstitute.org

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

Firehose Plays several key roles

- extreme scale production pipeline
- analytic forest-clearing for researchers & MDs
- democratization for use beyond TCGA proper
- simplification for everyone
- pushing envelope for rigor @ scale, reproducibility, APIs

At the height of sample characterization in TCGA, GDAC Firehose ingested 24K new data aliquots per year, with as many as 6K pipelines per month executed upon them

Rigorous pipelined analysis at unprecedented scale & complexity

2-3 orders of magnitude greater in scale & integration than leading cancer analyses in publication (circa 2011)

Data and analyses utilized at many academic, research and commercial sites around the world

Example: cBio@MSKCC

TCGA data & analyses in cBioPortal—expression, mutation, copy number, signficance analyses, and more—are loaded from GDAC Firehose.

More than 80K data aliquots from 11K cancer patients

But 

produces so much

In 2013 we published 62 Firehose runs

And today execute >1500 pipelines per analysis run

Compressing ~50TB of heterogeneous input to ~10GB results

Even this ~5000x distillation of data—>results can be daunting to wade through

Especially for individuals or small departments

Will only get bigger, faster & more integrated

NCI Genome Data Analysis Network (GDAN) — Fall 2016

- Will collect & analyze ~10K samples in 2 years
- This took TCGA 6-8 years

Firehose on Cloud: in ~public beta

- For entire research community
- More scalable & reproducible
- Streamline GDAN collaborations
- Rapid evolution of best-practice tools & workflows

firecloud.org

CPTAC: Clinical Proteomics Tumor Analysis Consortium

- 2016 RFA for Firehose-like proteo-genomic analysis center
- ~2K deeply characterized samples

FIREBROWSE

Search analysis results

HOME | BROAD GDAC | WEB API | TUTORIAL | RELEASE NOTES | ANALYSES GRAPH | FAQ | CONTACT

View Expression Profile | Enter gene name | Enter cohort abbrev | View Analysis Profile

SELECT COHORT

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- miR Analyses
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyses
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses

TCGA data version 2015_06_01

http://firebrowse.org

Simpler and more elegant way to explore

Sitting above one of the most comprehensive and deeply-characterized ***open*** cancer datasets in the world.

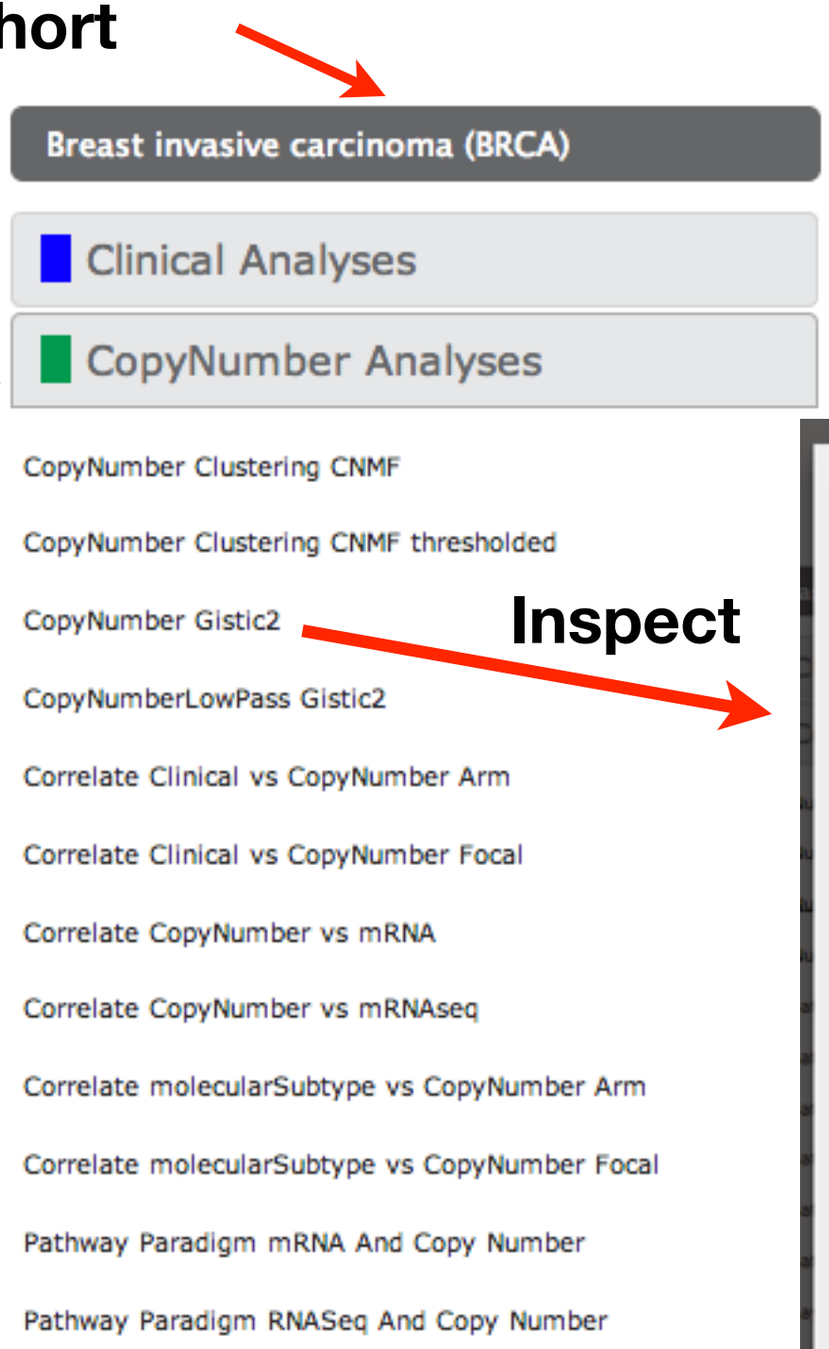
While retaining powerful GDAC Firehose backend, and offering advanced programmatic interfaces for experts

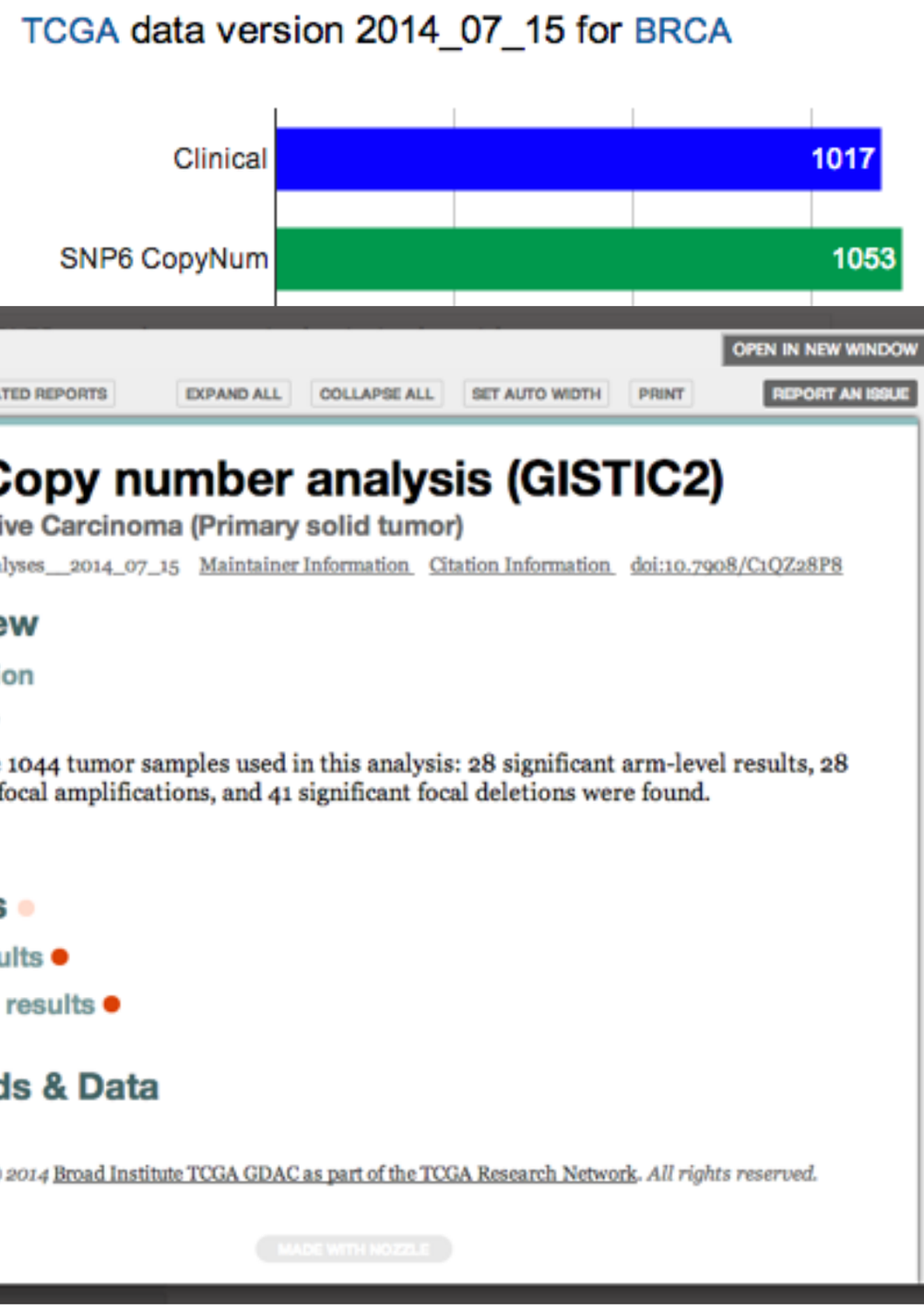
~1500 Analyses (reports) per run
Find your favorite in 2 clicks

Choose Cohort
(38 total)

Then DataType
(10 total)

Inspect

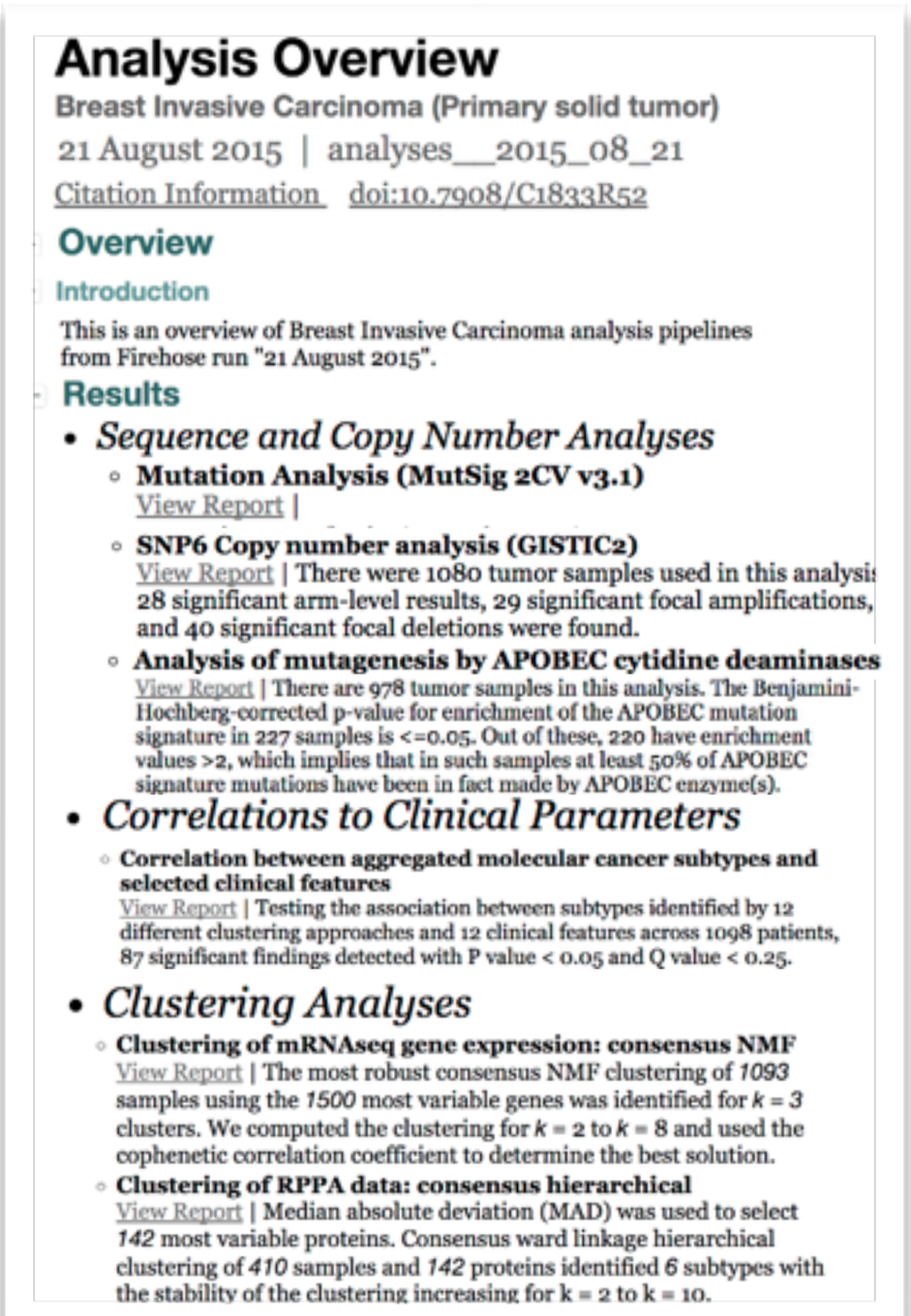




Organized like a paper

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

With Browser Convenience



Analysis Overview

Breast Invasive Carcinoma (Primary solid tumor)

21 August 2015 | analyses__2015_08_21

Citation Information doi:10.7908/C1833R52

Overview

Introduction

This is an overview of Breast Invasive Carcinoma analysis pipelines from Firehose run "21 August 2015".

Results

- *Sequence and Copy Number Analyses*
  - **Mutation Analysis (MutSig 2CV v3.1)**
    View Report |
  - **SNP6 Copy number analysis (GISTIC2)**
    View Report | There were 1080 tumor samples used in this analysis 28 significant arm-level results, 29 significant focal amplifications, and 40 significant focal deletions were found.
  - **Analysis of mutagenesis by APOBEC cytidine deaminases**
    View Report | There are 978 tumor samples in this analysis. The Benjamini-Hochberg-corrected p-value for enrichment of the APOBEC mutation signature in 227 samples is <=0.05. Out of these, 220 have enrichment values >2, which implies that in such samples at least 50% of APOBEC signature mutations have been in fact made by APOBEC enzyme(s).
- *Correlations to Clinical Parameters*
  - **Correlation between aggregated molecular cancer subtypes and selected clinical features**
    View Report | Testing the association between subtypes identified by 12 different clustering approaches and 12 clinical features across 1098 patients, 87 significant findings detected with P value < 0.05 and Q value < 0.25.
- *Clustering Analyses*
  - **Clustering of mRNAseq gene expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of *1093* 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 RPPA data: consensus hierarchical**
    View Report | Median absolute deviation (MAD) was used to select *142* most variable proteins. Consensus ward linkage hierarchical clustering of *410* samples and *142* proteins identified *6* subtypes with the stability of the clustering increasing for k = 2 to k = 10.

Organized like a paper

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

With Browser Convenience

Analysis Overview

Breast Invasive Carcinoma (Primary solid tumor)

21 August 2015 | analyses__2015_08_21

Citation Information doi:10.7908/C1833R52

Overview

Introduction

This is an overview of Breast Invasive Carcinoma analysis pipelines from Firehose run "21 August 2015".

Results

- *Sequence and Copy Number Analyses*
  - **Mutation Analysis (MutSig 2CV v3.1)**
    View Report |
  - **SNP6 Copy number analysis (GISTIC2)**
    View Report | There were 1080 tumor samples used in this analysis: 28 significant arm-level results, 29 significant focal amplifications, and 40 significant focal deletions were found.
  - **Analysis of mutagenesis by APOBEC cytidine deaminases**
    View Report | There are 978 tumor samples in this analysis. The Benjamini-Hochberg-corrected p-value for enrichment of the APOBEC mutation signature in 227 samples is <=0.05. Out of these, 220 have enrichment values >2, which implies that in such samples at least 50% of APOBEC signature mutations have been in fact made by APOBEC enzyme(s).
- *Correlations to Clinical Parameters*
  - **Correlation between aggregated molecular cancer subtypes and selected clinical features**
    View Report | Testing the association between subtypes identified by 12 different clustering approaches and 12 clinical features across 1098 patients, 87 significant findings detected with P value < 0.05 and Q value < 0.25.
- *Clustering Analyses*
  - **Clustering of mRNAseq gene expression: consensus NMF**
    View Report | The most robust consensus NMF clustering of *1093* 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 RPPA data: consensus hierarchical**
    View Report | Median absolute deviation (MAD) was used to select *142* most variable proteins. Consensus ward linkage hierarchical clustering of *410* samples and *142* proteins identified *6* subtypes with the stability of the clustering increasing for k = 2 to k = 10.

SNP6 Copy number analysis (GISTIC2)

Breast Invasive Carcinoma (Primary solid tumor)

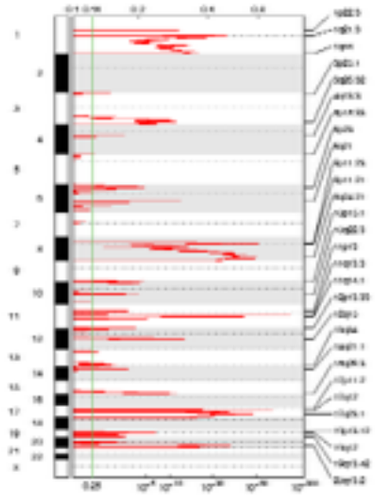
21 August 2015 | analyses__2015_08_21 Maintainer Information Citation Information

Summary

There were 1080 tumor samples used in this analysis: 28 significant arm-level results, 29 significant focal amplifications, and 40 significant focal deletions were found.

Results

Focal results



Cytoband	Q value	Residual Q value	Wide Peak Boundaries	# Genes in Wide Peak
11q13.3	4.0078e-213	2.7106e-192	chr11:69412882-69487994	2
8q24.21	8.6006e-92	8.6006e-92	chr8:128676088-128770551	1
8p11.23	1.7328e-96	2.1059e-83	chr8:37492669-37604543	2
17q12	8.5344e-137	1.2297e-70	chr17:37790163-37876887	6

Clustering of mRNAseq gene expression

Breast Invasive Carcinoma (Primary solid tumor)

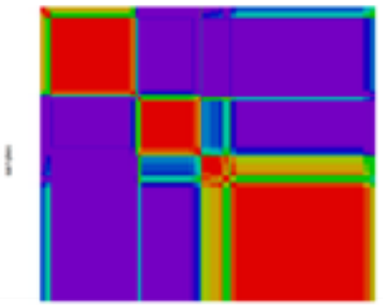
21 August 2015 | analyses__2015_08_21 Maintainer Information Citation Information

Summary

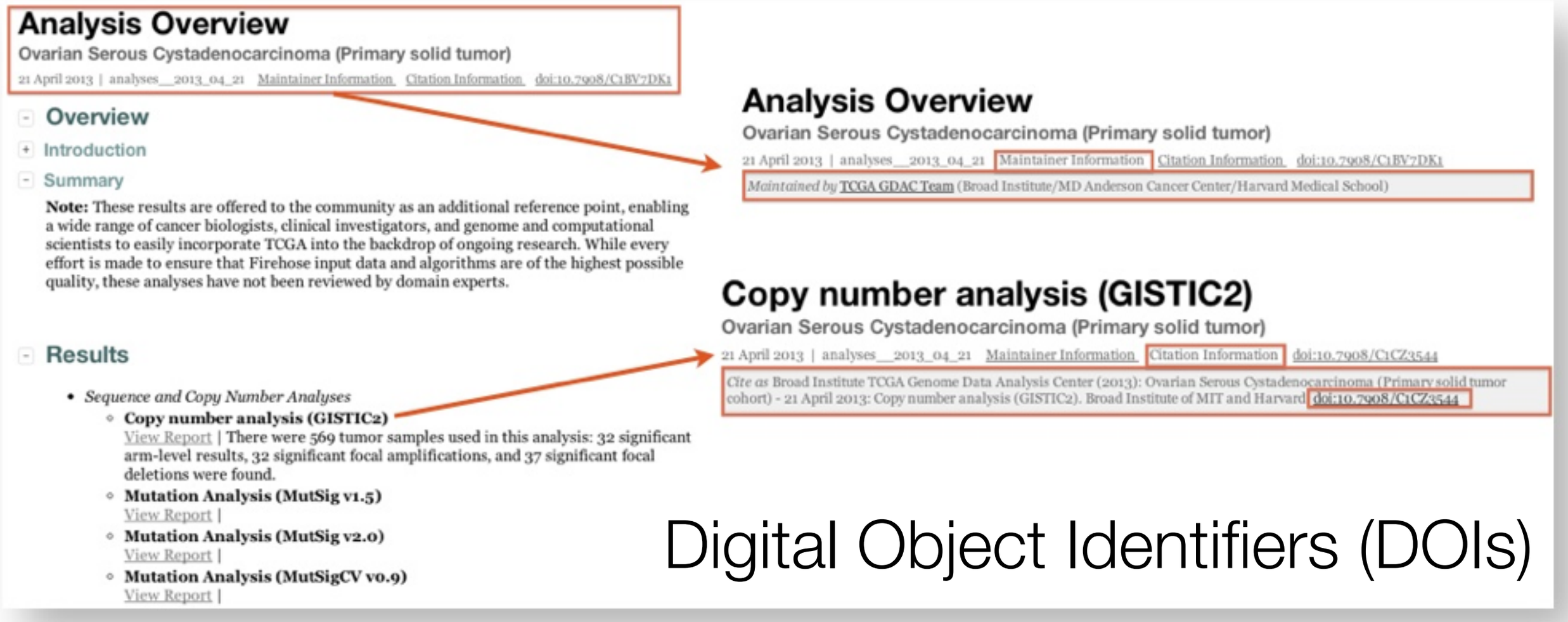
The most robust consensus NMF clustering of *1093* samples using the *1500* most variable genes was identified for *k* = *3* clusters. We computed the clustering for *k* = 2 to *k* = 8 and used the cophenetic correlation coefficient to determine the best solution.

Consensus and correlation matrix

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



Directly Citable in The Literature



Digital Object Identifiers (DOIs)

Akin to creating 1500 draft manuscripts per run

Many 1000s of datasets per run
Find your favorite in 2 clicks

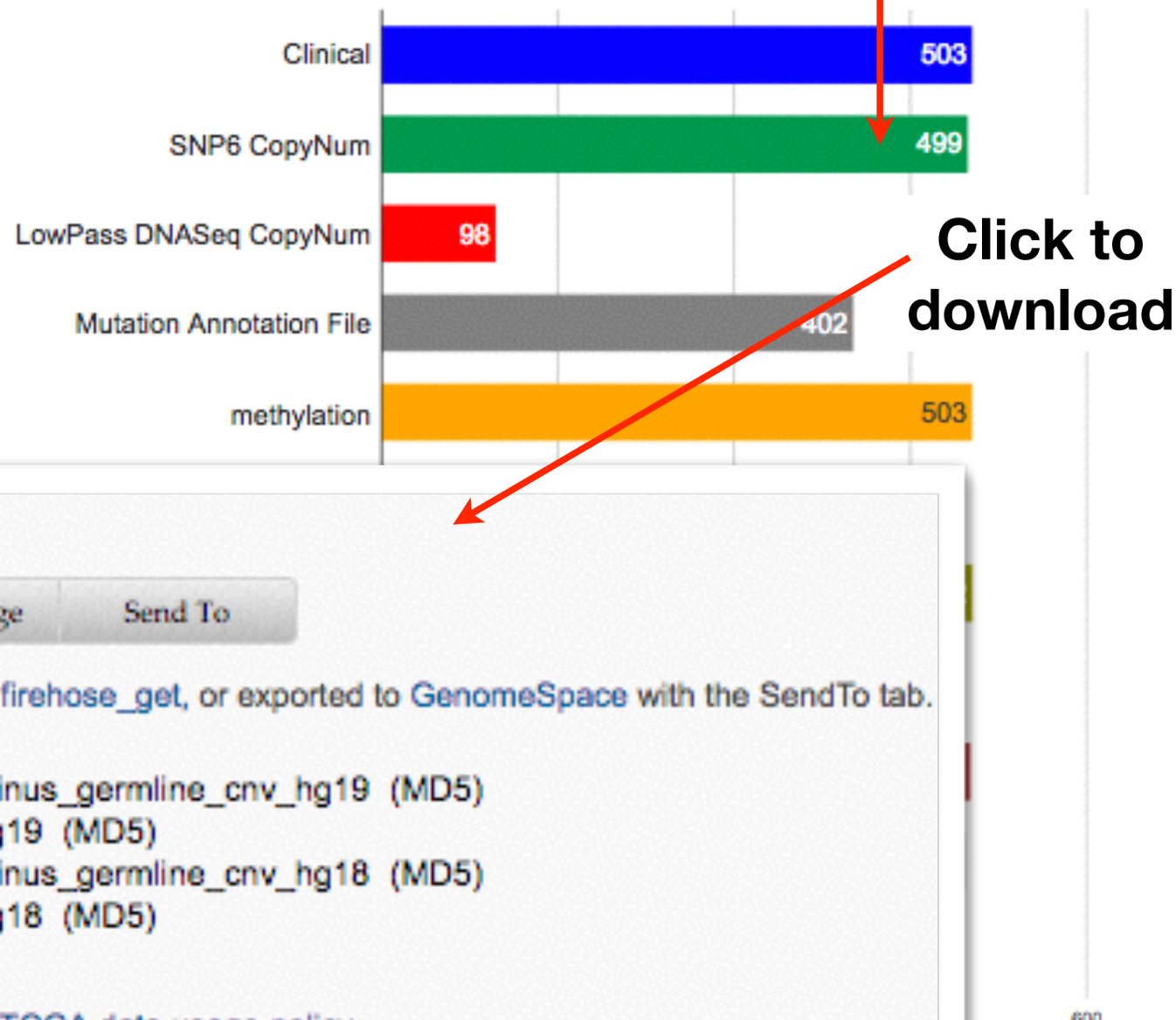
Choose Cohort

Then DataType

Thyroid carcinoma (THCA)

- Clinical Analyses
- CopyNumber Analyses
- Correlations Analyses
- Methylation Analyses
- miRseq Analyses
- mRNA An
- mRNAseq
- Mutation
- Pathway
- RPPA Ana

TCGA data version 2016_01_28 for THCA



Click to download

THCA CopyNumber Archives

Primary | Auxiliary | SDRF/Mage | Send To

Files may also be downloaded here, or with firehose_get, or exported to GenomeSpace with the SendTo tab.

genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg19 (MD5)
genome_wide_snp_6-segmented_scna_hg19 (MD5)
genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg18 (MD5)
genome_wide_snp_6-segmented_scna_hg18 (MD5)

Downloading data constitutes agreement to TCGA data usage policy

Or easily send to GenomeSpace for more analysis

THCA CopyNumber Archives

Primary | Auxiliary | SDRF/Mage | Send To

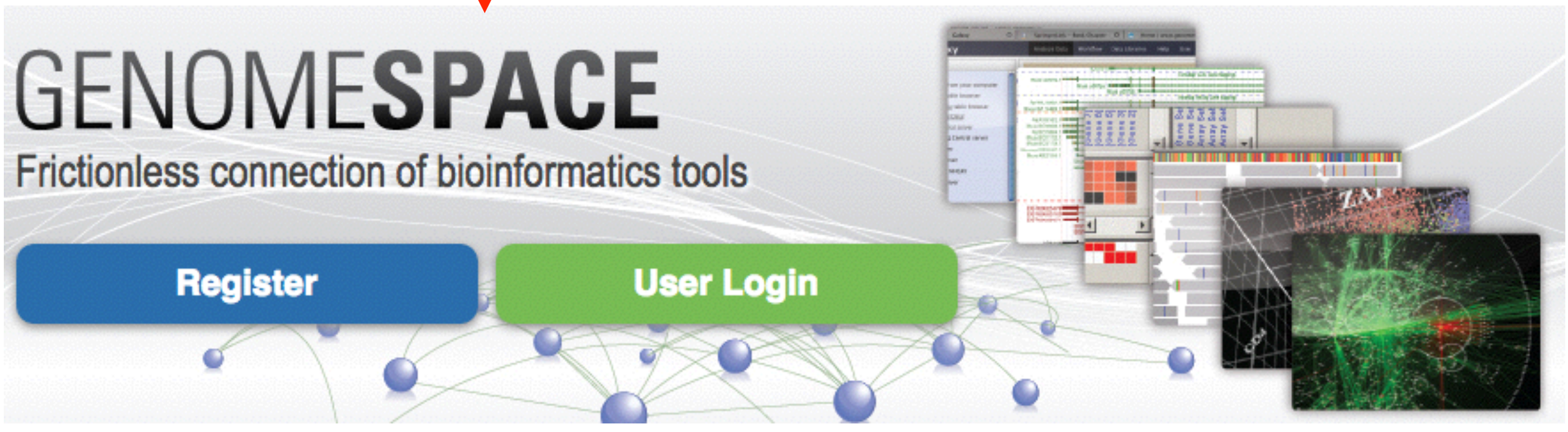
Files may also be downloaded here, or with firehose_get, or exported to GenomeSpace with the SendTo tab.

- [x] **genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg19 [851.18 KB]**
- [ ] genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg18 [852.54 KB]

Cumulative file size for current selections: 851.18 KB

GENOMESPACE BETA Upload | Clear Selections

Downloading data constitutes agreement to TCGA data usage policy



Or download everything with 1 command

```
linux%  firehose_get  analyses  latest
```

```
linux%  firehose_get    data    latest
```

Simple 20K BASH script, just 1 moving part

Obtain Here

Programmatic Tools*

* Crafted to reasonate with biomedical researchers more than SWEs

FireBrowse UI powered by 25+ RESTful apis in 4 categories

HOME | BROAD GDAC | WEB API | ANALYSES GRAPH | FAQ | CONTACT

Analyses : Fine grained retrieval of analysis pipeline results — Show/Hide | List Operations | Expand Operations | Raw

- GET /Analyses/Mutation/MAF — Retrieve MutSig final analysis MAF.
- GET /Analyses/Mutation/SMG — Retrieve Significantly Mutated Genes (SMG).
- GET /Analyses/CopyNumber/Genes/All
- GET /Analyses/CopyNumber/Genes/Focal
- GET /Analyses/CopyNumber/Genes/Thresholded
- GET /Analyses/CopyNumber/Genes/Amplified
- GET /Analyses/CopyNumber/Genes/Deleted
- 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

Retrieve GisticZ significantly amplified genes results.

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

Show/Hide | List Operations

- GET /Archives/StandardData

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

Show/Hide | List Operations | Expa

- GET /Metadata/Counts
- GET /Metadata/Cohorts — Retrieve map of cohort abbreviation
- GET /Metadata/Cohort/{cohort} — Retrie
- GET /Metadata/Platforms — Retrieve map of platform code(s

Interactive Docs

Don't Need to be a Programmer

GET /Samples/mRNASeq — Retrieve mRNASeq data.

Implementation Notes

This service returns sample-level log2 mRNASeq expression values. Results may be filtered by gene, cohort, barcode, sample type or characterization protocol, but at least one gene OR barcode must be supplied.

Parameters

Parameter	Value	Description	Parameter Type	Data Type
format	json (default)	Format of result.	query	string
gene	egfr	Comma separated list of gene name(s).	query	string
cohort	ACC BLCA BRCA CESC	Narrow search to one or more TCGA disease cohorts from the scrollable list.	query	string
tcga_participant_barcode		Comma separated list of TCGA participant barcodes (e.g. TCGA-GF-A4EO).	query	string
sample_type	NB NT TAM TAP	Narrow search to one or more TCGA sample types from the scrollable list.	query	string
protocol	RPKM RSEM	Narrow search to one or more sample characterization protocols from the scrollable list.	query	string

Interactive Docs

Don't Need to be a Programmer

Explore data by playing in real time
instead of cut/paste from static HTML or PDF
Great way to learn APIs

GET /Samples/mRNASeq — Retrieve mRNASeq data.

Implementation Notes

This service returns sample-level log2 mRNASeq expression values. Results may be filtered by gene, cohort, barcode, sample type or characterization protocol, but at least one gene OR barcode must be supplied.

Parameters

Parameter	Value	Description	Parameter Type	Data Type
format	json (default)	Format of result.	query	string
gene	egfr	Comma separated list of gene name(s).	query	string
cohort	ACC BLCA BRCA CESC	Narrow search to one or more TCGA disease cohorts from the scrollable list.	query	string
tcga_participant_barcode		Comma separated list of TCGA participant barcodes (e.g. TCGA-GF-A4EO).	query	string
sample_type	NB NT TAM TAP	Narrow search to one or more TCGA sample types from the scrollable list.	query	string
protocol	RPKM RSEM	Narrow search to one or more sample characterization protocols from the scrollable list.	query	string

Interactive Docs

Don't Need to be a Programmer

Explore data by playing in real time
instead of cut/paste from static HTML or PDF
Great way to learn APIs

GET /Samples/mRNASeq — Retrieve mRNASeq data.

Implementation Notes

This service returns sample-level log2 mRNASeq expression values. Results may be filtered by gene, cohort, barcode, sample type or characterization protocol, but at least one gene OR barcode must be supplied.

Parameters

Parameter	Value	Description	Parameter Type	Data Type
format	json (default)	Format of result.	query	string
gene	egfr	Comma separated list of gene name(s).	query	string
cohort	ACC BLCA BRCA CESC	Narrow search to one or more TCGA disease cohorts from the scrollable list.	query	string
tcga_participant_barcode		Comma separated list of TCGA participant	query	string
sample_type	NB NT TAM TAP	from the scrollable list.		
protocol	RPKM RSEM	Narrow search to one or more sample characterization protocols from the scrollable list.	query	string

choices clearly enumerated

automatically generated & updated as API and database evolve

Perform Query Hide Response

Request URL

```
http://firebrowse.org:8000/api/v1/Samples/mRNASeq?format=json&gene=egfr&page=1&page_size=250&sort_by=gene
```

Proper RESTful call is ASSEMBLED FOR YOU


```
{
  "cohort": "ACC",
  "expression_log2": 7.59666610237019,
  "gene": "EGFR",
  "geneID": 1956,
  "protocol": "RSEM",
  "sample_type": "TP",
  "tcga_participant_barcode": "TCGA-OR-A5J1",
  "z-score": -0.40056053472322
},
{
  "cohort": "ACC",
  "expression_log2": 6.98214823852598,
  "gene": "EGFR",
  "geneID": 1956,
  "protocol": "RSEM",
  "sample_type": "TP",
  "tcga_participant_barcode": "TCGA-OR-A5J2",
  "z-score": -0.572210443678677
},
```


JSON for computers/programmers

Results returned in multiple formats

tcga_participant_barcode	gene	expression_log2	z-score	cohort	sample_type	
TCGA-OR-A5J1	EGFR	7.59666610237	-0.400560534723	ACC	TP	RSEM
TCGA-OR-A5J2	EGFR	6.98214823853	-0.572210443679	ACC	TP	RSEM
TCGA-OR-A5J3	EGFR	9.31231960446	0.729969055244	ACC	TP	RSEM
TCGA-OR-A5J5	EGFR	8.50495520815	0.0333590221281	ACC	TP	RSEM
TCGA-OR-A5J6	EGFR	8.5592941021	0.0690092698339	ACC	TP	RSEM
TCGA-OR-A5J7	EGFR	8.64932911891	0.131115969294	ACC	TP	RSEM
TCGA-OR-A5J8	EGFR	8.06454015357	-0.210987070006	ACC	TP	RSEM
TCGA-OR-A5J9	EGFR	6.63334692474	-0.641628460792	ACC	TP	RSEM
TCGA-OR-A5JA	EGFR	9.05879837786	0.468028706825	ACC	TP	RSEM
TCGA-OR-A5JB	EGFR	8.50794128032	0.0352834298625	ACC	TP	RSEM
TCGA-OR-A5JC	EGFR	7.55685241318	-0.414030877529	ACC	TP	RSEM
TCGA-OR-A5JD	EGFR	6.25656347946	-0.699966368647	ACC	TP	RSEM
TCGA-OR-A5JE	EGFR	6.16656683008	-0.711787657396	ACC	TP	RSEM
TCGA-OR-A5JF	EGFR	8.56235233966	0.0710558865356	ACC	TP	RSEM
TCGA-OR-A5JG	EGFR	8.96827107766	0.385101741143	ACC	TP	RSEM
TCGA-OR-A5JI	EGFR	7.05755857856	-0.554865718674	ACC	TP	RSEM
TCGA-OR-A5JJ	EGFR	6.64321260426	-0.639886855174	ACC	TP	RSEM

TSV, CSV for scientists, algorithms

Python and UNIX CLI Bindings

- *Automatically generated from interactive docs*
- BSD-style open source
- Install with PyPI or obtain from FireBrowse
- Extensively documented: Python, CLI, R, WWW

Docs for virtually all class methods & functions can ***also*** be obtained by invoking the function with zero arguments

(better than several inscrutable pages of a stack trace, don't you think?)

FireBrowseR : R bindings

- Mario Deng et al, Ph.D. candidate @ Bonn
- Available on GitHub (and soon CRAN)

Powerful but simple queries: EGFR expression

```
linux%   fbget  mrnaseq  egfr  cohort=ucs
```

tcga_participant_barcode	gene	expression_log2	z-score	cohort
TCGA-QN-A5NN	EGFR	7.06162500905	-0.59899352506	UCS
TCGA-QM-A5NM	EGFR	8.16734387649	-0.29844359375	UCS
TCGA-NG-A4VW	EGFR	8.93092623547	0.09326678880	UCS

Because @ times even writing a few lines of Python takes too long

Powerful but simple queries: EGFR expression

```
linux%    fbget  mrnaseq  egfr
```

tcga_participant_barcode	gene	expression_log2	z-score	cohort
TCGA-QN-A5NN	EGFR	7.06162500905	-0.59899352506	UCS
TCGA-QM-A5NM	EGFR	8.16734387649	-0.29844359375	UCS
TCGA-NG-A4VW	EGFR	8.93092623547	0.09326678880	UCS

Because @ times even writing a few lines of Python takes too long

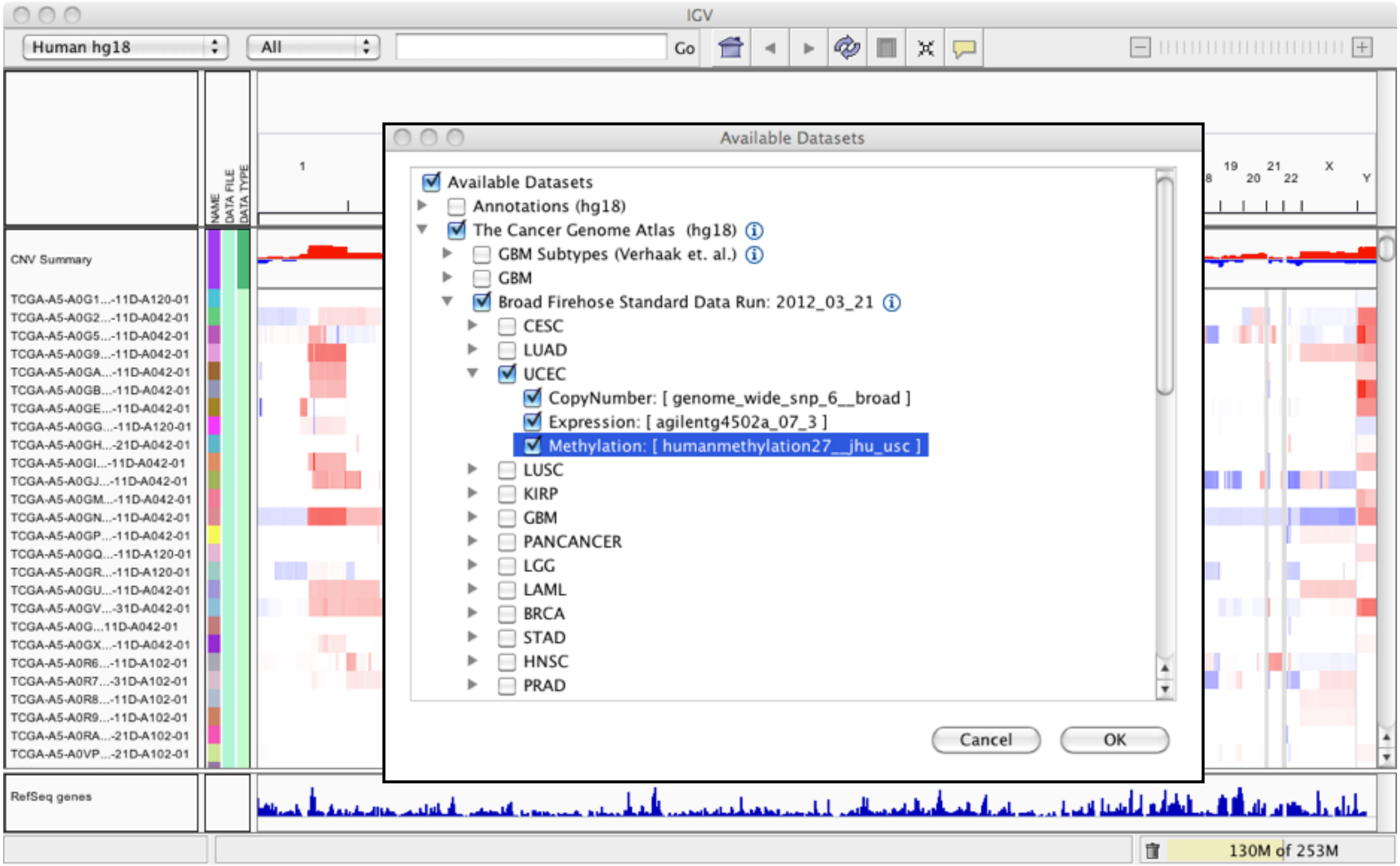
Coarse or fine grained

Get all samples in a single cohort
Or for ALL patients in TCGA
Or even a single patient

just omit cohort

Graphical Tools

Data pre-loaded into IGV (for years)



If you have IGV you have FireBrowse data

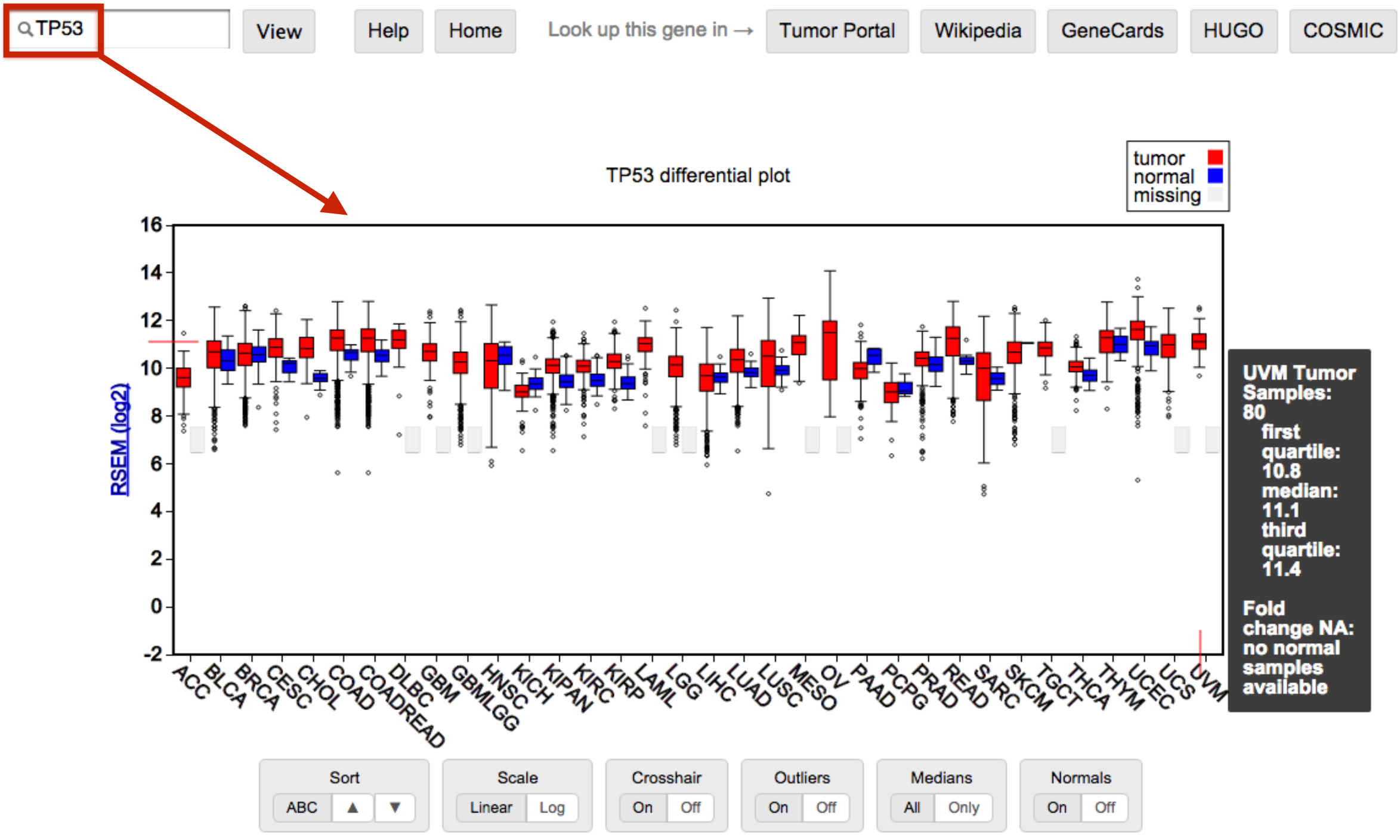
Easily send mutation, expression & CN data to

ArrayExpress | cBioPortal | CCLE | Cistrome | Cytoscape | Cytoscape 3 | FireBrowse | Galaxy | GenePattern

Extends analytic reach of FireBrowse with cloud-based workbench, for easy data flow in chains of many interactive analysis tools

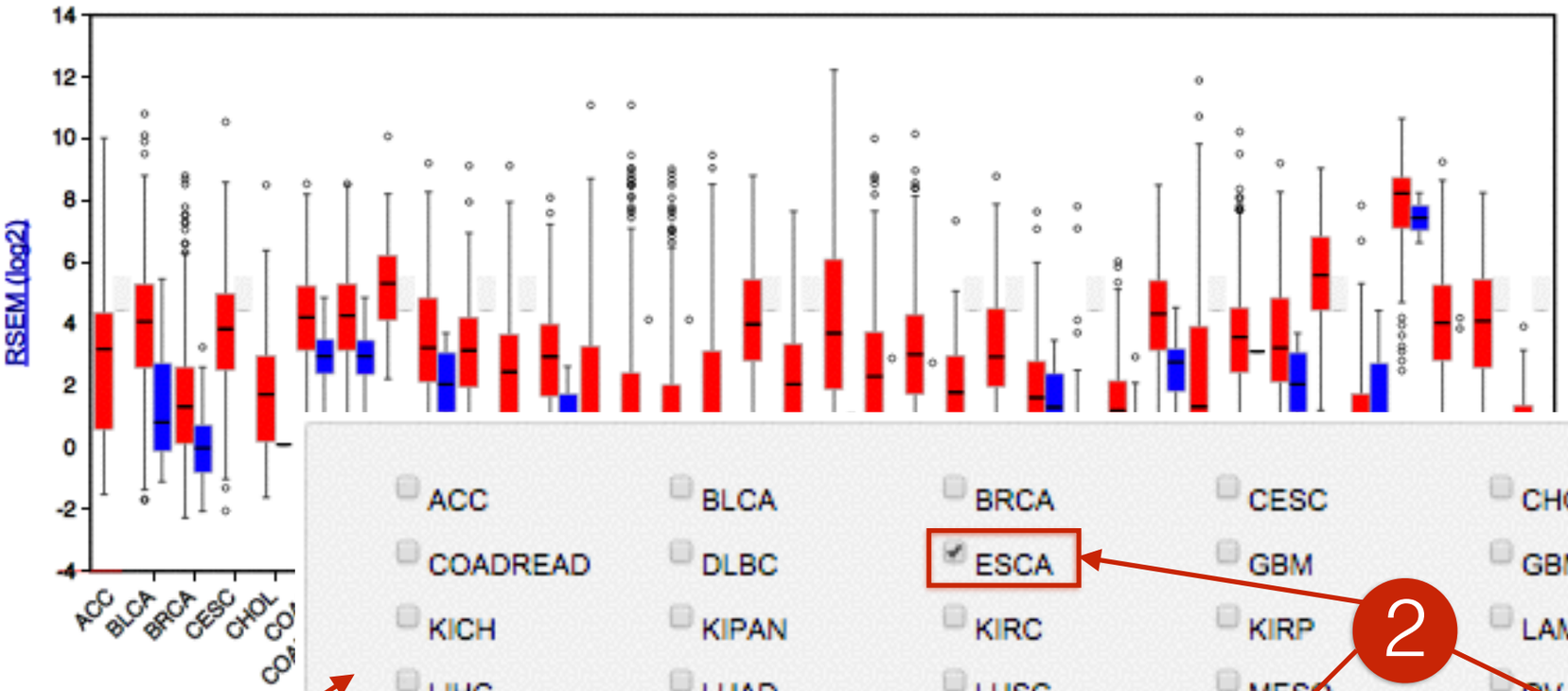
More data types will be exportable soon

viewGene: expression level browser

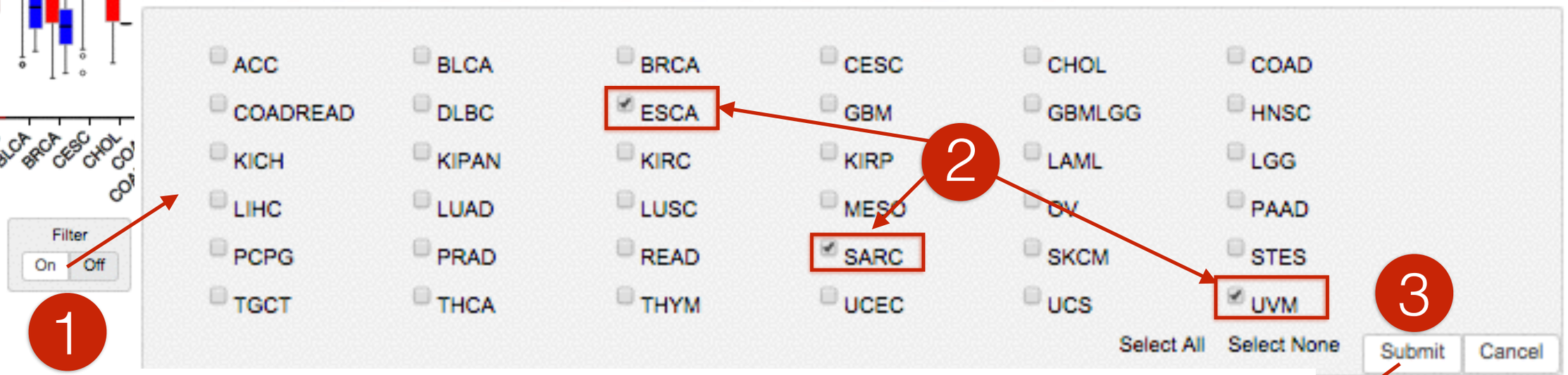


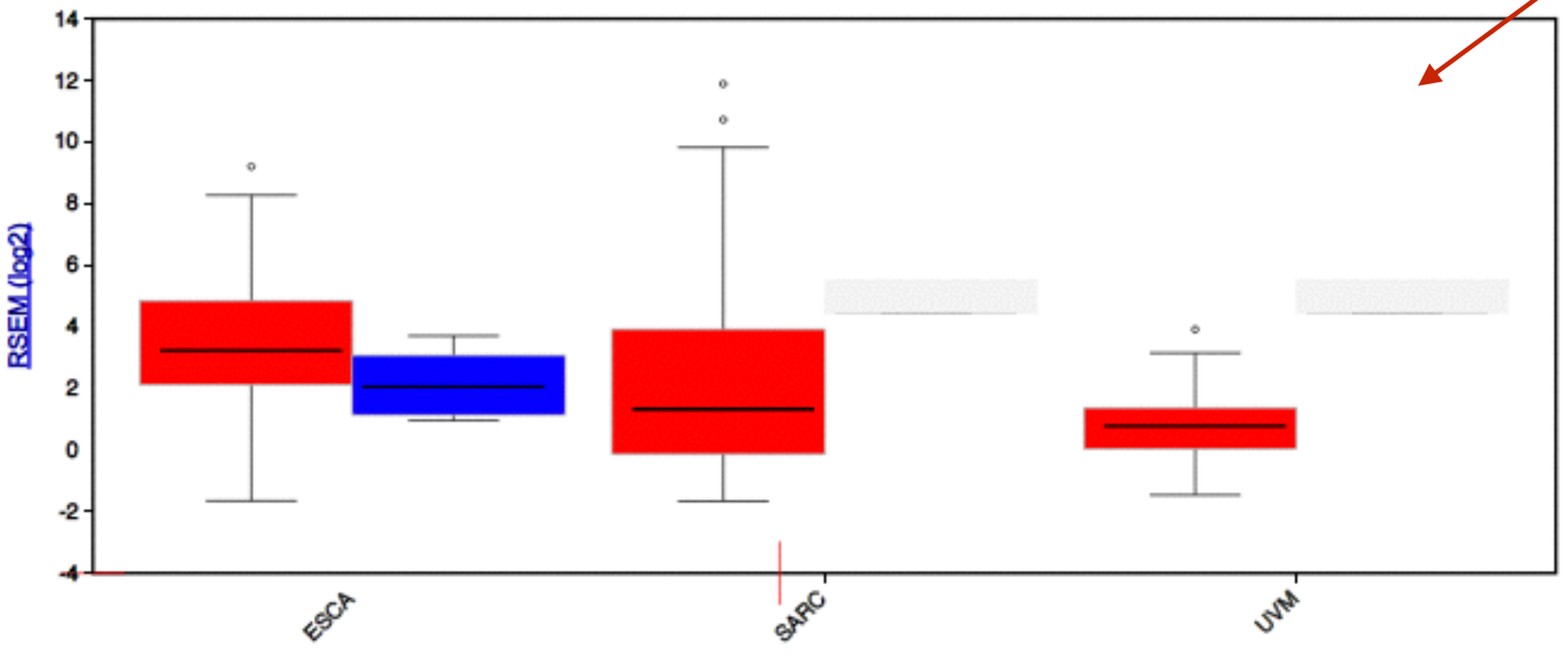
Backed by FireBrowse API, quickly inspect gene expression levels

TERT differential plot

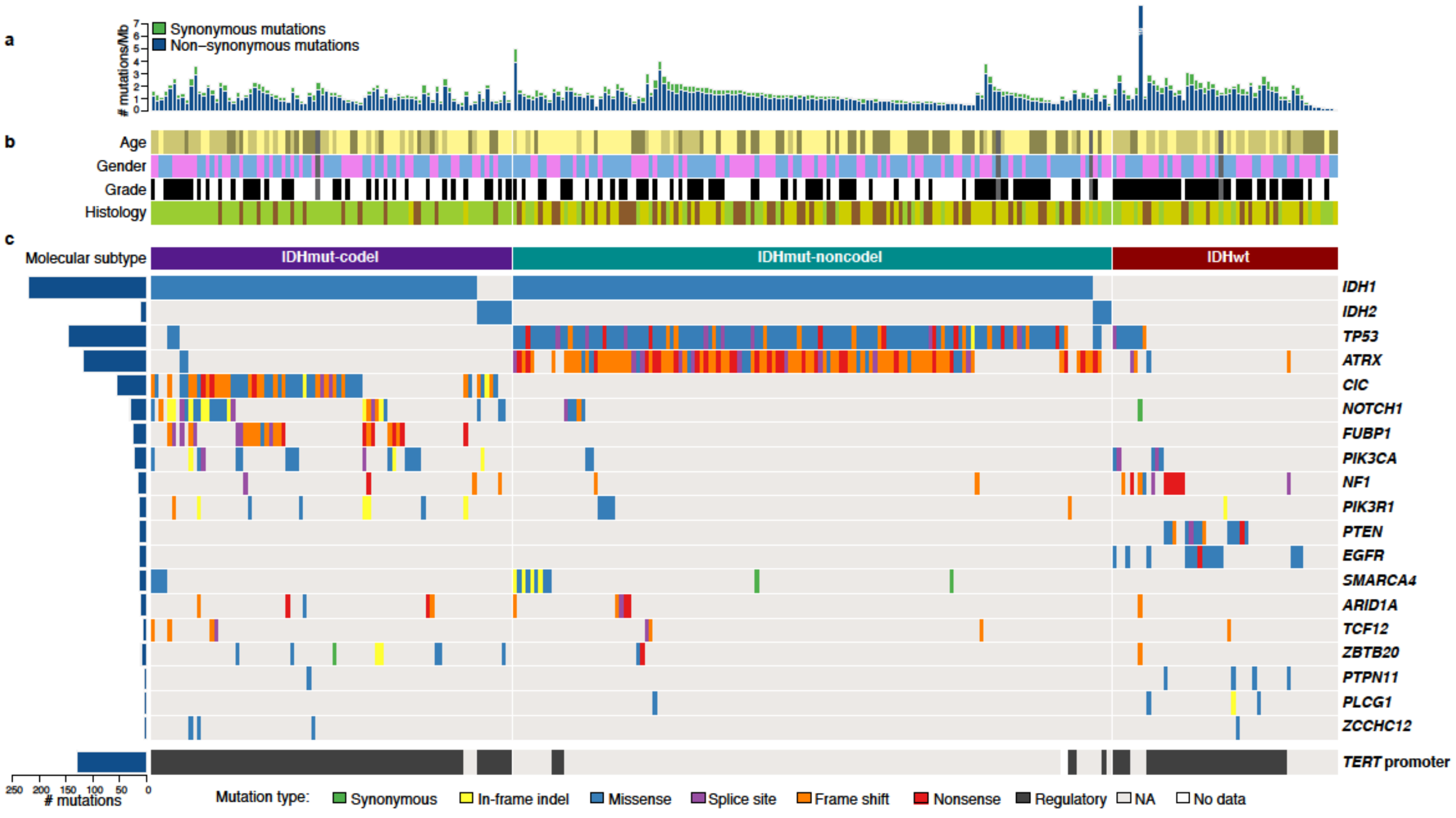


View expression levels across all cohorts, or arbitrary subsets.





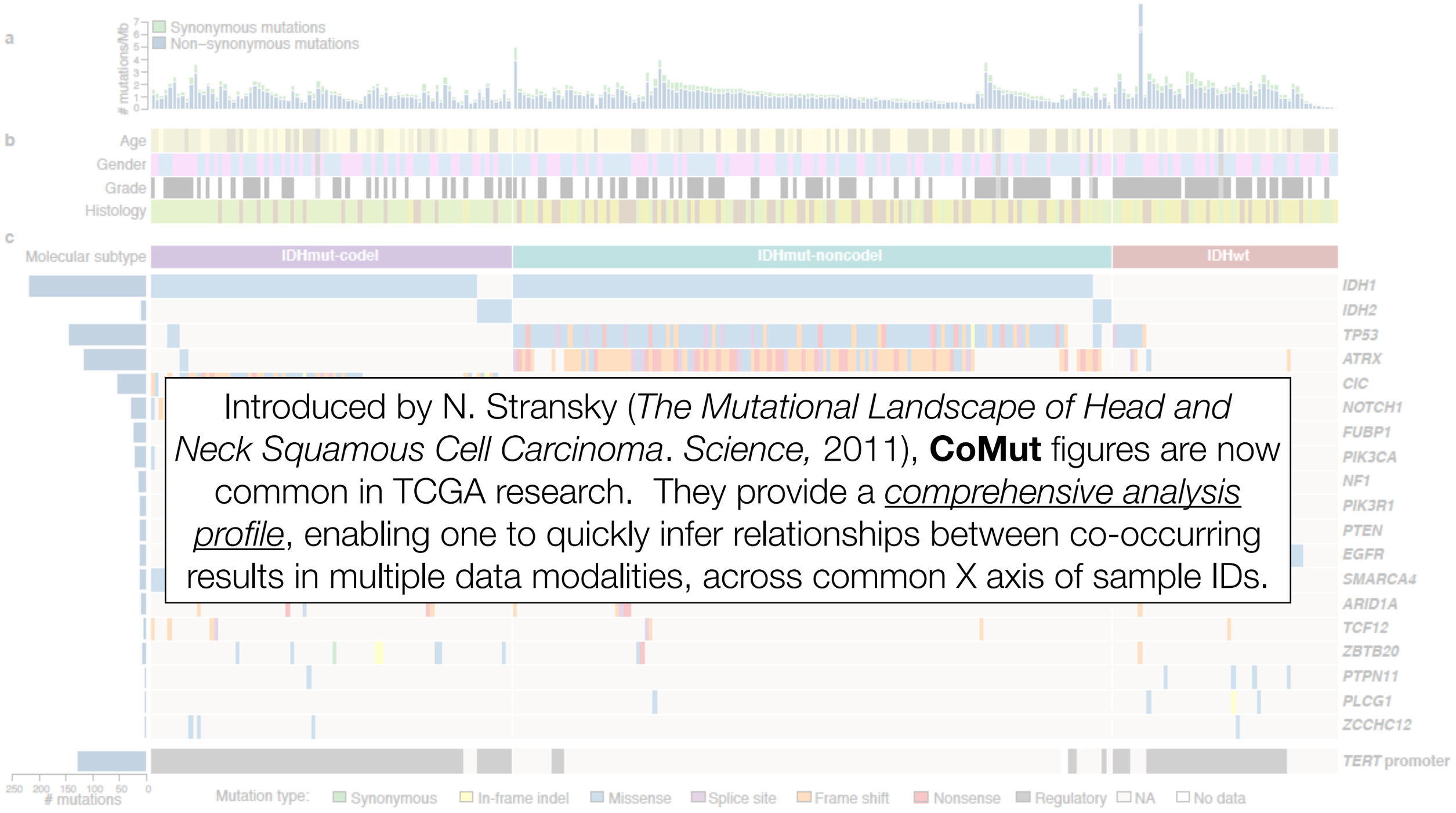
But our most exciting new tool compressess an entire Firehose run into a single, interactive & reproducible figure



Comprehensive and Integrative Genomic Characterization of Diffuse Lower Grade Gliomas (TCGA Network 2015)

Jaegil Kim, Broad Institute

But our most exciting new tool compressess an entire Firehose run into a single, interactive & reproducible figure



Introduced by N. Stransky (*The Mutational Landscape of Head and Neck Squamous Cell Carcinoma*. *Science,* 2011), **CoMut** figures are now common in TCGA research. They provide a *comprehensive analysis profile*, enabling one to quickly infer relationships between co-occurring results in multiple data modalities, across common X axis of sample IDs.

Comprehensive and Integrative Genomic Characterization of Diffuse Lower Grade Gliomas (TCGA Network 2015)

Jaegil Kim, Broad Institute

But in journals, figures are static and can be small and hard to read

And cannot be explored in real time

And reproducing them or investigating their implications can require substantial time for data retrieval, preparation and analysis

iCoMut

By making such figures interactive, allowing panels to be moved, sorted and searched, iCoMut dramatically enhances that process.

Astrocytoma

Oligoastrocytoma

Oligodendroglioma

Mutation Rate

synonymous

non synonymous

Mutation rate

Gene Mutation
x: TCGA-IK-7675
y: IDH1
Missense

Clinical parameters

age

vital status

gender

histology

Show Grids Hide Grids

IDH1
IDH2
TP53
ATRX
CIC
NOTCH1
FUBP1
PIK3CA
NF1
ARID1A
ZBTB20
TCF12
PTPN11
PDGFRA*
ZCCHC12

NA
Nonsense
Frameshift

Mutation significance

No Mutation

Mutations

Focal Level CN Gain

CN gain

Loss
Deletion
No Change

Amp_10p15.3
Amp_7p11.2
Amp_13q34
Amp_12q14.1
Amp_4q12
Amp_7q32.3
Amp_8q24.13
Amp_12p13.32
Amp_7q31.2
Amp_Xp11.22
Amp_1q32.1
Amp_19p13.3
Amp_11q24.1
Amp_3q26.33
Amp_17q25.1

#SCNAs

Focal Level CN Loss

CN loss

NA
Amplification
Gain
Loss
Deletion
No Change

Del_19q13.42
Del_19q13.41
Del_1p36.32
Del_1p32.3
Del_9p21.3
Del_9p23
Del_10q26.3
Del_14q24.3
Del_11p15.5
Del_13q14.2
Del_6q24.3
Del_4q34.3
Del_6q22.31
Del_11p15.1
Del_22q13.31
Del_18q23
Del_2q37.3
Del_6p25.3
Del_12q12
Del_5p15.33
Del_3p21.31
Del_3q29
Del_5q34
Del_Xq21.1

#SCNAs

CLUS_mRNA_cNMF

CLUS_mRNA_cHierarchical

Clusters

RPPA cHierarchical

mRNAseq cNMF

mRNAseq cHierarchical

miRseq cNMF

miRseq cHierarchical

miRseq Mature cNMF

miRseq Mature cHierarchical

mRNA cNMF
mRNA cHierarchical
CN cNMF
Methylation cNMF
RPPA cNMF
RPPA cHierarchical
mRNAseq cNMF
mRNAseq cHierarchical
miRseq cNMF
miRseq cHierarchical
miRseq Mature cNMF
miRseq Mature cHierarchical

iCoMut

By making such figures <u>interactive,</u> allowing panels to be moved, sorted and searched, iCoMut dramatically enhances that process.

Example: hovering over pixels tells you about the underlying biology.

iCoMut

By making such figures interactive, allowing panels to be moved, sorted and searched, iCoMut dramatically enhances that process.

Example: hovering over pixels tells you about the underlying biology.

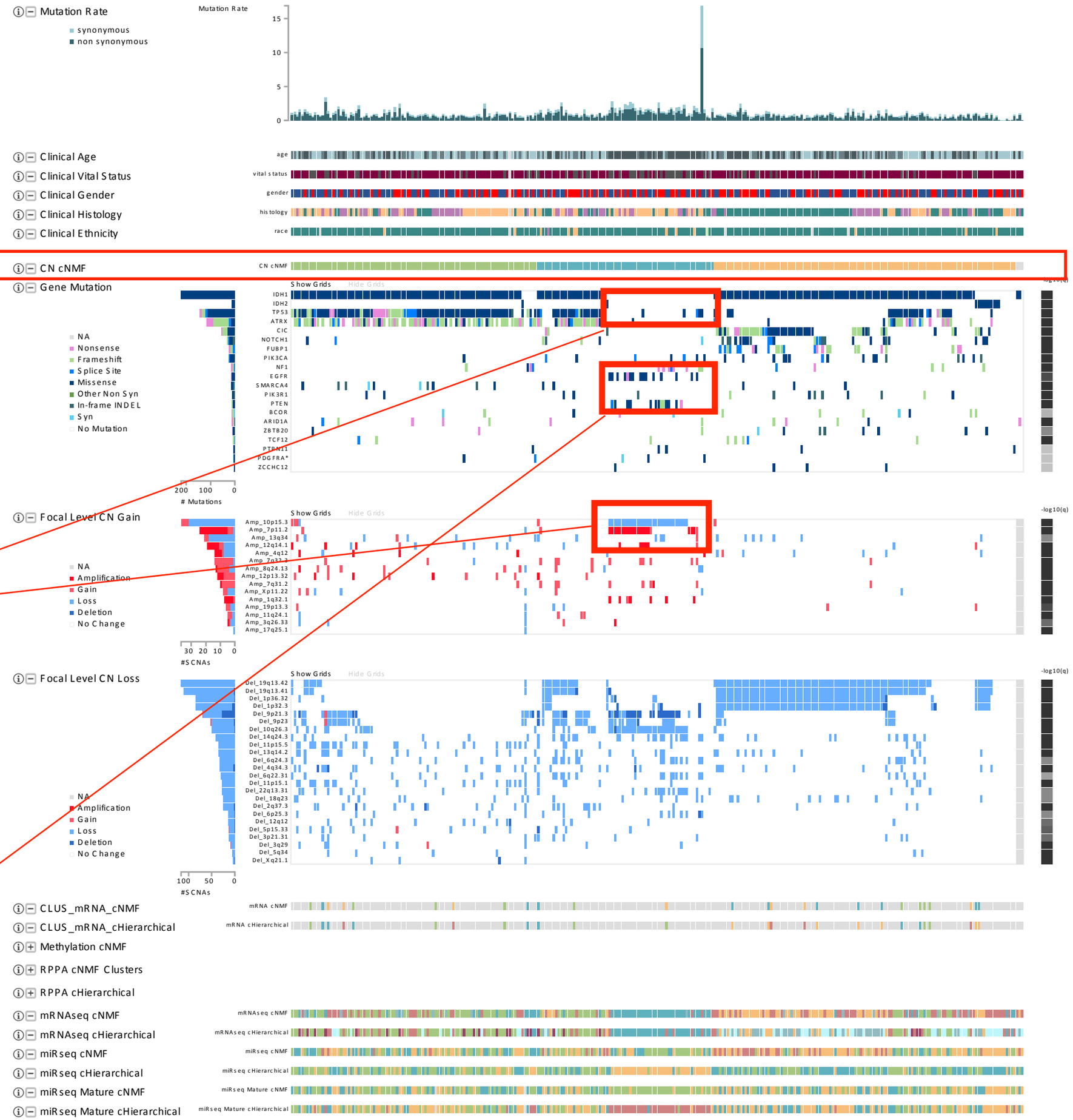
Here we show the TCGA LGG cohort: sorted first by clinical histology, then gene (descending order of mutation count). The clinical subtypes leap off the page at you.

As does the fact that the copy-number landscape differs when IDH1/2, TP53, and ATRX mutations drop off.

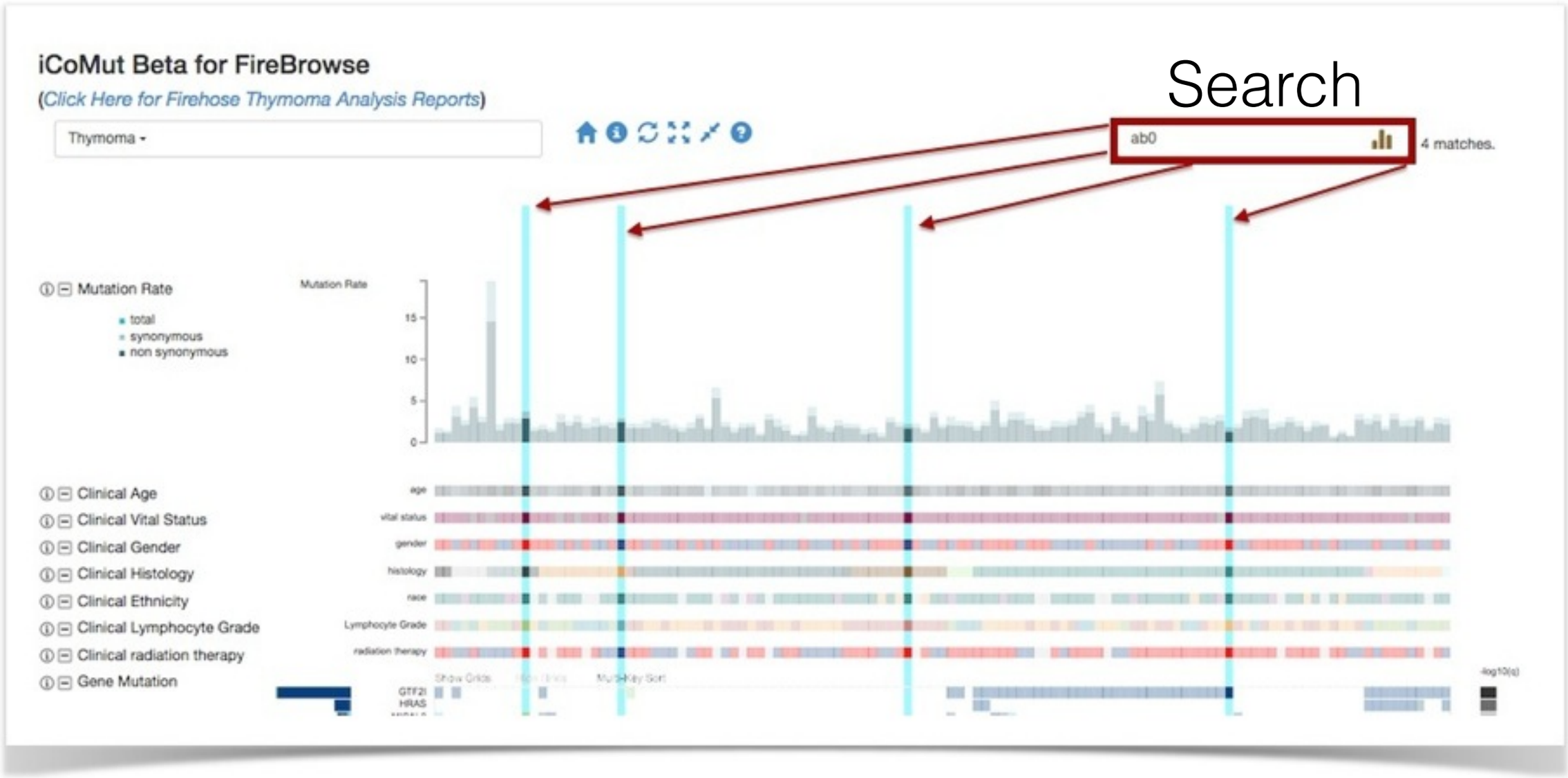
Now we've re-sorted by CNMF copy-number clustering, ***and dragged it from bottom of figure to top,*** just above mutation panel

Making it further apparent that the copy-number landscape differs as IDH1/2, TP53, and ATRX mutations diminish

Also shows apparent involvement with EGFR and PTEN.

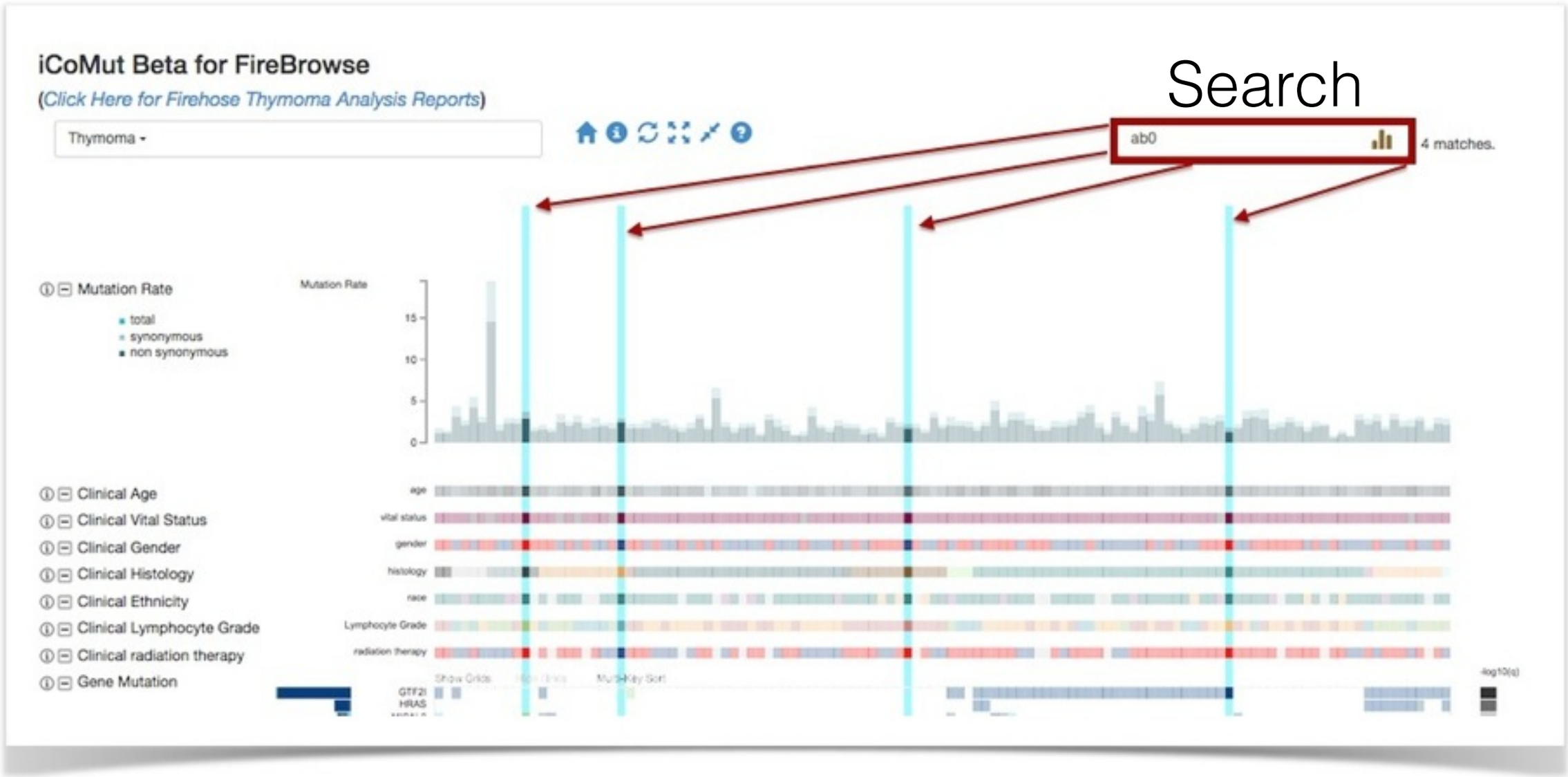


Many more graphical controls …



Example: locate patient/sample of interest

Many more graphical controls …



Example: locate patient/sample of interest

Collaboratively explore questions in realtime on telecons: in what expression cluster does patient X fall?

Without database lookup or scripting, etc

Advanced Search

Include these samples: OR-A5K5

Exclude these samples: C5-A0TN

Panel Functions

	Row	OP	Value
mutation_rate	non_synonymous	>	

-- Select a Panel --
✓ mutation_rate
clinical_age
clinical_vital_status
clinical_gender
clinical_histology
clinical_ethnicity
gene_mutation
focal_level_cn_gain
focal_level_cn_loss
mrnaseq_cnmf
mrnaseq_chierarchical
mirseq_cnmf
mirseq_cHierarchical
mirseq_mature_cnmf
mirseq_mature_chierarchical
cn_cnmf
clus_methylation_cnmf
rppa_cnmf_clusters
rppa_chierarchical

✓ non_synonymous
synonymous
total

✓ >
>=
<
<=
=
!=

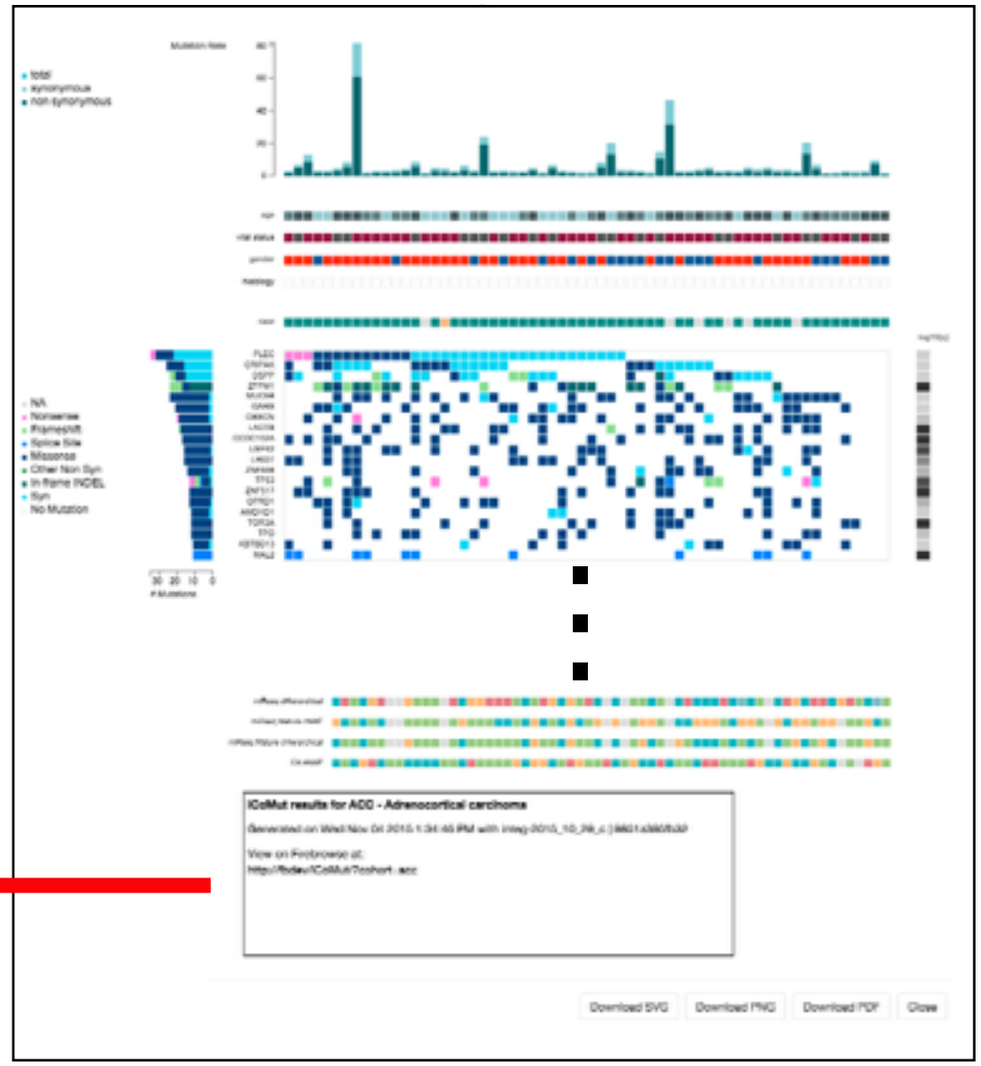
AND OR

Search

Hide Grids Multi-Key Sort

Push-button publication figure reproducibility

SVG



iCoMut results for ACC - Adrenocortical carcinoma

Generated on Wed Nov 04 2015 1:34:46 PM with integ-2015_10_29_c | 9851a395fb32

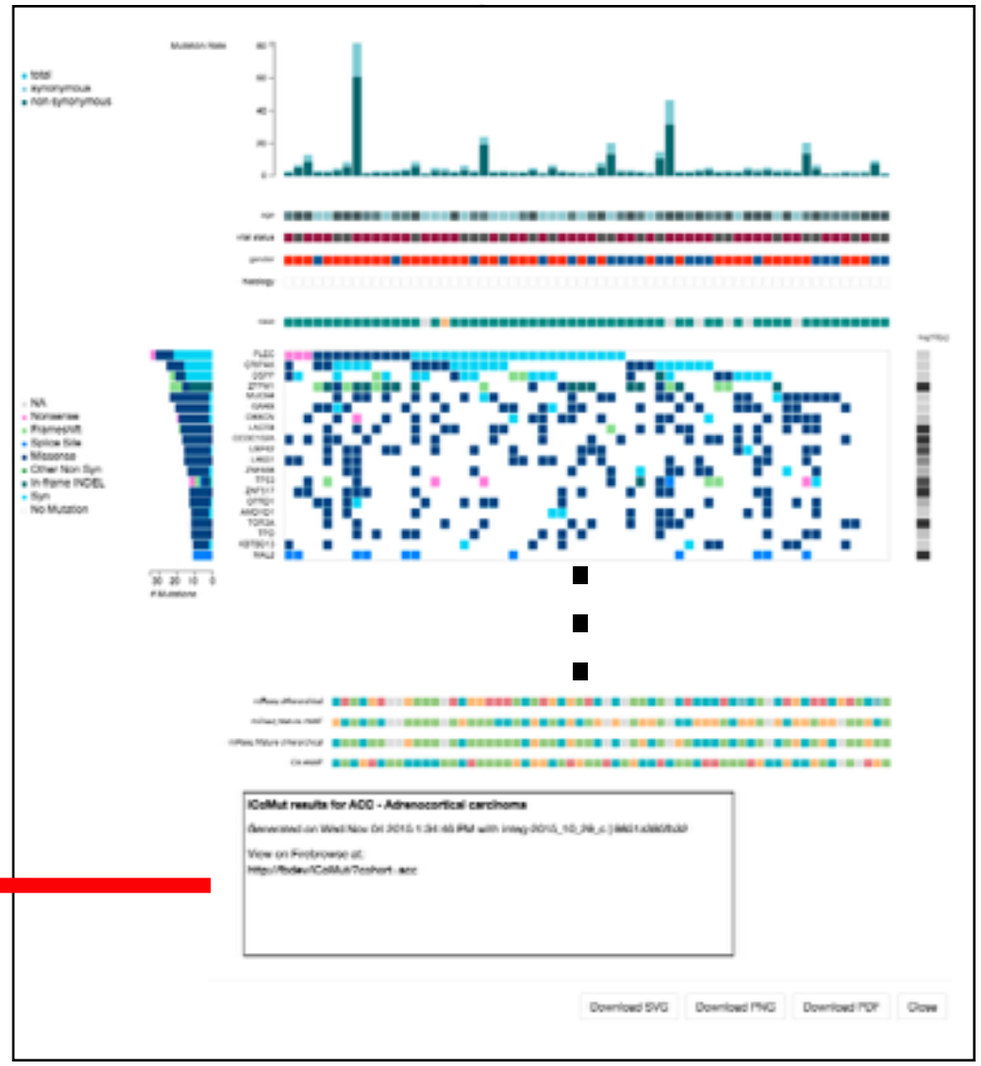
View on Firebrowse at:

http://firebrowse.org/iCoMut/?cohort=acc

URL to regenerate: reflecting the interactive manipulations to figure

Push-button publication figure reproducibility

SVG



iCoMut results for ACC - Adrenocortical ca

Generated on Wed Nov 04 2015 1:34:46 PM

View on Firebrowse at:

http://firebrowse.org/iCoMut/?cohort=acc

Much more useful than simple screenshot

URL to regenerate: reflecting the interactive manipulations to figure

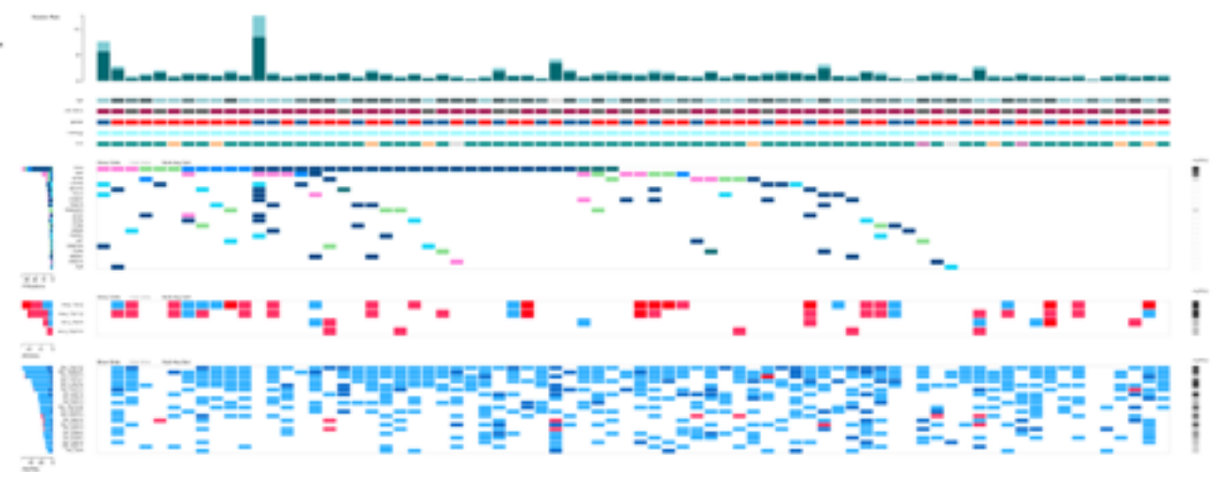
Growing Momentum

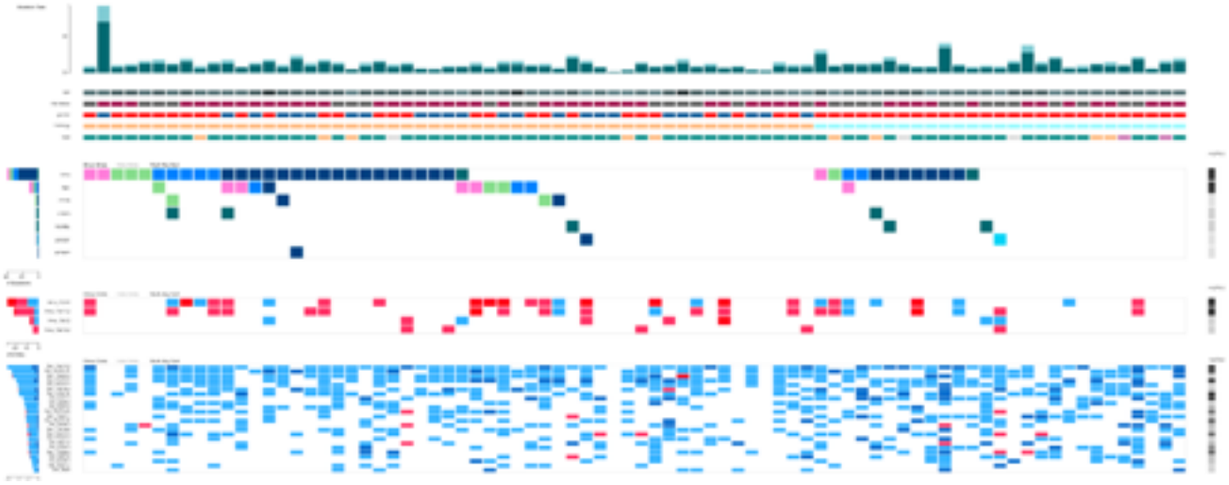
(~10 months old)

- Regularly use in TCGA analysis working groups (AWGs)
  - ✓ Encapsulates 7 TCGA awg runs to date
  - ✓ 2-3 more AWG runs within ~1 month
  - ✓ Summary figure for recent series of TCGA AWG papers
  - ✓ Cited in additional papers too

- Adopted for 4 portals already: FireBrowse, and
  - ✓ tumorportal.org
  - ✓ Mass General Hospital: internal clinical portal
  - ✓ Multiple Myeloma Foundation: MMPG portal (soon)

Bonus: visual diff tool





Sarcoma AWG LMS subtype
2015_07_30

2016_02_10

iCoMut makes it easy to spot changes between runs
Despite extremely high complexity & information density

Where Next?

- Semi-open: GitHub by invitation to early collaborators
- Planning to be fully open later this year
- Next release version ~1 month:
  - JSON input : more flexible for custom data
  - More crisp / high performing drag-n-drop

- Further plans for 2016:
  - More APIs : methylation, protein, correlations
  - Rubberband selection for data export
  - More visual features
  - Deploy for more TCGA AWGs / manuscripts
  - Ramp up for use in GDAN (or “TCGA 2.0”)
  - Science-oriented front-end for FireCloud workspaces

Pushing to Raise the Bar

For simplicity & accessibility in the presentation of information-dense, high-throughput science

Couched in memes that resonate with scientists

Enabling BOTH experts AND researchers with little or no TCGA, bioinformatics or programming experience

To leap to the forefront of cancer research with just a few clicks on their desktop

FIREBROWSE

Fin

Acknowledgements

PI: **Gad Getz,** Lynda Chin

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

David Heiman	Hailei Zhang
Katherine Huang	Juok Cho
Kane Hadley	Jaegil Kim
Sam Meier	Tim DeFreitas

The front line computational biologists
and software engineers.

Alumni: D. DiCara, H. Arachchi, B. Alexander, W. Mallard, R. Zupko, R. Sinha