Resources for Web-based exploration of multi-point Cancer Genomics Datasets

TCGA Analysis Workshop
Jan 7, 2016

Parthav Jailwala
CCR Collaborative Bioinformatics Resource (CCBR)

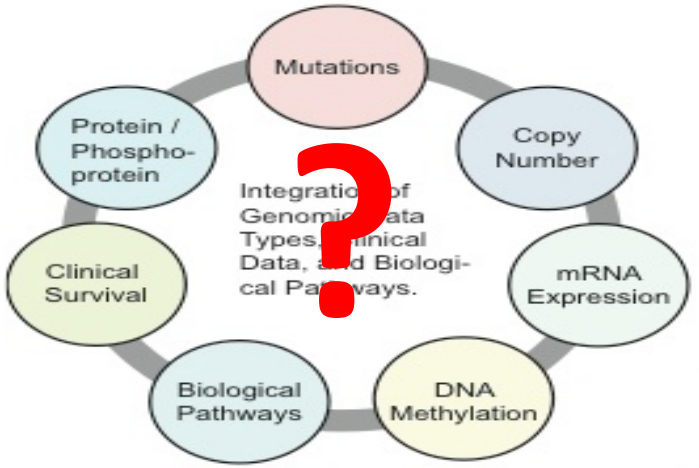
I know gene X is highly expressed in the cancer type I am researching; Is it also highly expressed in other closely related cancers?

What cluster of genes are highly correlated in expression, between cancer X and Y?

Is the mutation in gene X and gene Y mutually exclusive?

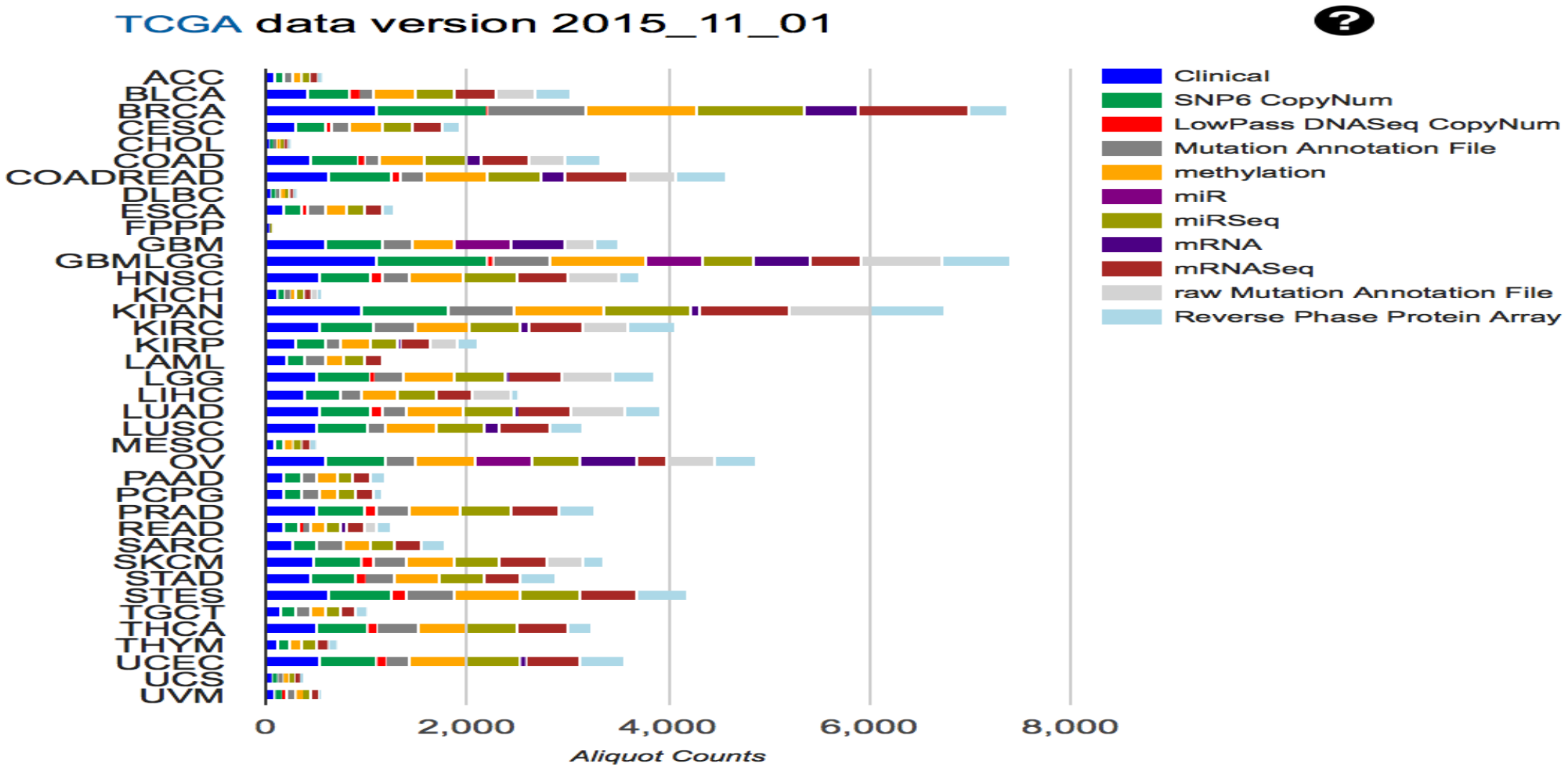
Large-scale Cancer Genomics Studies (TCGA, ICGC)

Petabytes of multi-dimensional information from thousands of patients



Integrative multi-point analyses & biological interpretation

TCGA cohorts and data types



Data collection & characterisation

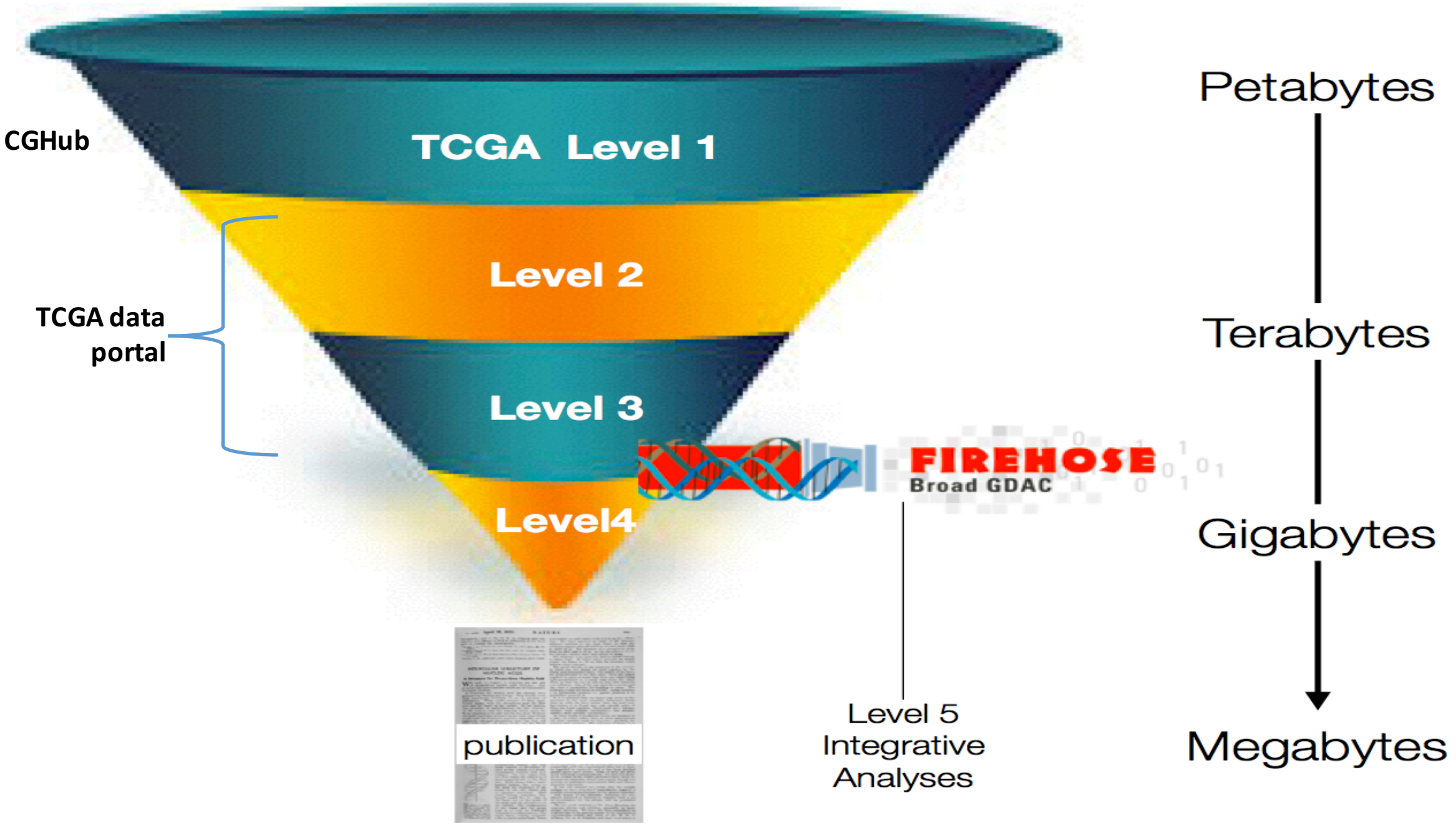
Data storage & integrative analyses

Reports, data visualization & biological data interpretation

The Cancer Genome Atlas

FIREHOSE
Broad GDAC

FIREBROWSE : Mining the Firehose of TCGA


CGHub
TCGA Level 1
Level 2
Level 3
Level4
TCGA data portal
FIREHOSE
Broad GDAC
publication
Level 5 Integrative Analyses
Petabytes
Terabytes
Gigabytes
Megabytes

Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Browse
Bladder urothelial carcinoma	BLCA	412	Browse	Browse
Breast invasive carcinoma	BRCA	1098	Browse	Browse
Cervical and endocervical cancers	CESC	307	Browse	Browse
Cholangiocarcinoma	CHOL	36	Browse	Browse
Colon adenocarcinoma	COAD	460	Browse	Browse
Colorectal adenocarcinoma	COADREAD	631	Browse	Browse
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browse	Browse
Esophageal carcinoma	ESCA	185	Browse	Browse
FFPE Pilot Phase II	FPPP	38	None	Browse
Glioblastoma multiforme	GBM	613	Browse	Browse
Glioma	GBMLGG	1129	Browse	Browse
Head and Neck squamous cell carcinoma	HNSC	528	Browse	Browse
Kidney Chromophobe	KICH	113	Browse	Browse
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	Browse	Browse
Kidney renal clear cell carcinoma	KIRC	537	Browse	Browse
Kidney renal papillary cell carcinoma	KIRP	323	Browse	Browse
Acute Myeloid Leukemia	LAML	200	Browse	Browse
Brain Lower Grade Glioma	LGG	516	Browse	Browse
Liver hepatocellular carcinoma	LIHC	377	Browse	Browse
Lung adenocarcinoma	LUAD	585	Browse	Browse
Lung squamous cell carcinoma	LUSC	504	Browse	Browse
Mesothelioma	MESO	87	Browse	Browse
Ovarian serous cystadenocarcinoma	OV	602	Browse	Browse
Pancreatic adenocarcinoma	PAAD	185	Browse	Browse
Pheochromocytoma and Paraganglioma	PCPG	179	Browse	Browse
Prostate adenocarcinoma	PRAD	499	Browse	Browse
Rectum adenocarcinoma	READ	171	Browse	Browse
Sarcoma	SARC	260	Browse	Browse
Skin Cutaneous Melanoma	SKCM	470	Browse	Browse
Stomach adenocarcinoma	STAD	443	Browse	Browse
Stomach and Esophageal carcinoma	STES	628	Browse	Browse
Testicular Germ Cell Tumors	TGCT	150	Browse	Browse
Thyroid carcinoma	THCA	503	Browse	Browse

http://gdac.broadinstitute.org

38 disease cohorts

~80K aliquots

~1500 result reports per analysis run

Cite-able with DOIs

Completely open

Every aliquot described in detailed samples report

Millions of hits across world

The GDAC FireHose offers…

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

2. Version-stamped package of standard analyses results
   - Automatically generated for dozens of algorithms: GISTIC, MutSig, Clustering, Correlation, …

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**

Three modes of integrative data mining

- **Discovery mode / Hypothesis generation**
  - Finding ***global correlations*** across heterogeneous datatypes

FireBrowse: iCoMut
Next generation clustered heatmaps (NG-CHM)
UCSC Cancer Genomics Browser -> Xena Browser
BioDiscovery NexusDB

→ **Create NG-CHMs for your own multi-point data**

- **Confirmation mode / Explain mechanism of already observed correlations**
  - Finding correlations for a ***few genes/genomic regions*** across heterogeneous datatypes

cBioPortal
FireBrowse: ViewGene
canEvolve Web Portal (Dana Farber Institute)
PROGgeneV2: Pan-cancer prognostics database

→ **Create OncoPrints for your own multi-point data**

A simple and elegant way to explore cancer data.

Backed by a powerful computational infrastructure, application programming interface (API), graphical tools and online reports.

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

With over 80K sample aliquots from 11,000+ cancer patients, spanning 38 unique disease cohorts.

FireBrowse: Simplified Portal Access

Easy to find any of the TCGA datasets or Firehose analysis result reports

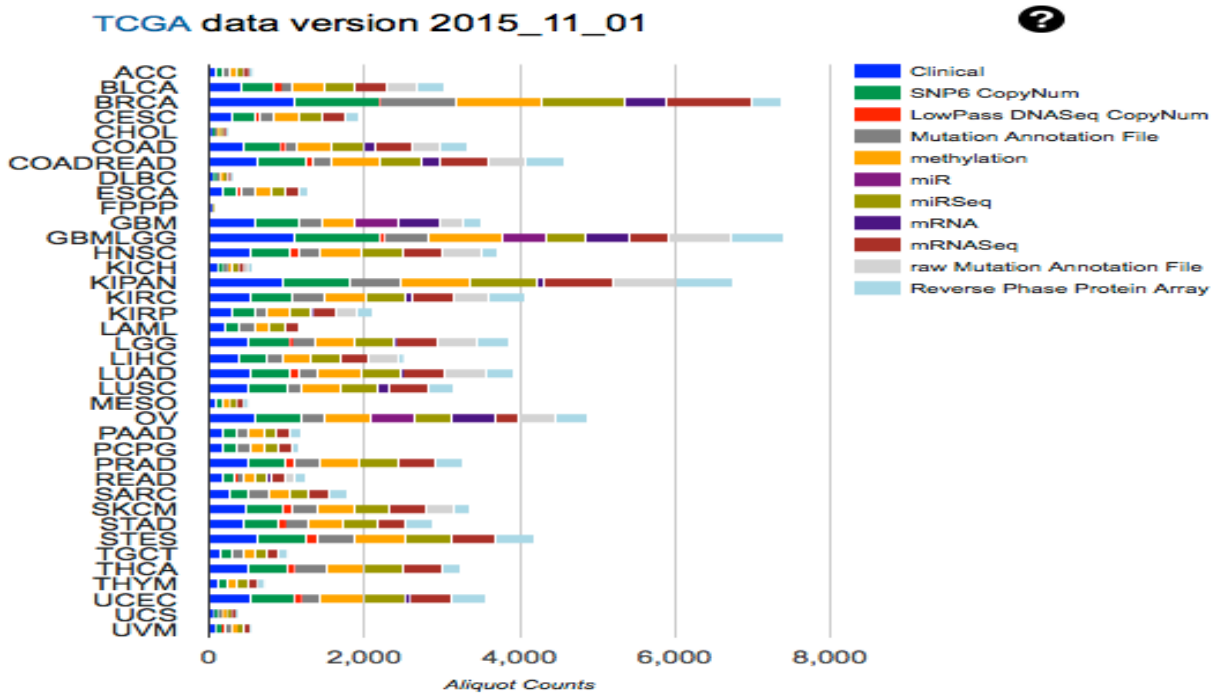
FIREBROWSE beta

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

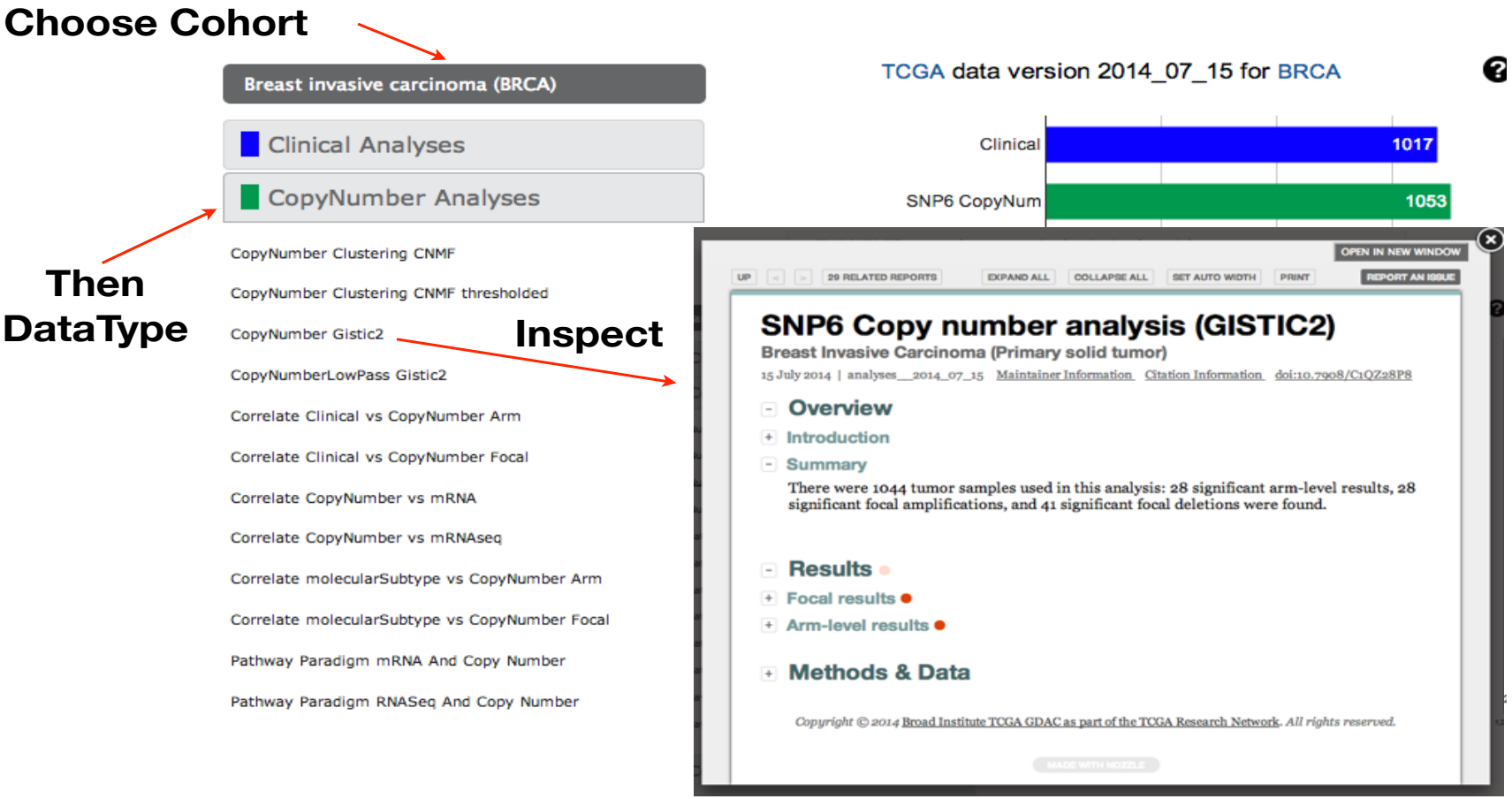


What can you do using FireBrowse ?

1. Access analysis reports for a chosen cancer cohort + datatype
2. Access version stamped datasets to generate your own analysis + reports
3. Visualize expression profile of one gene across cancer cohorts
4. Visualize integrative multi-panel analysis profiles for a chosen cancer cohort
5. Construct your own interactive queries using the API

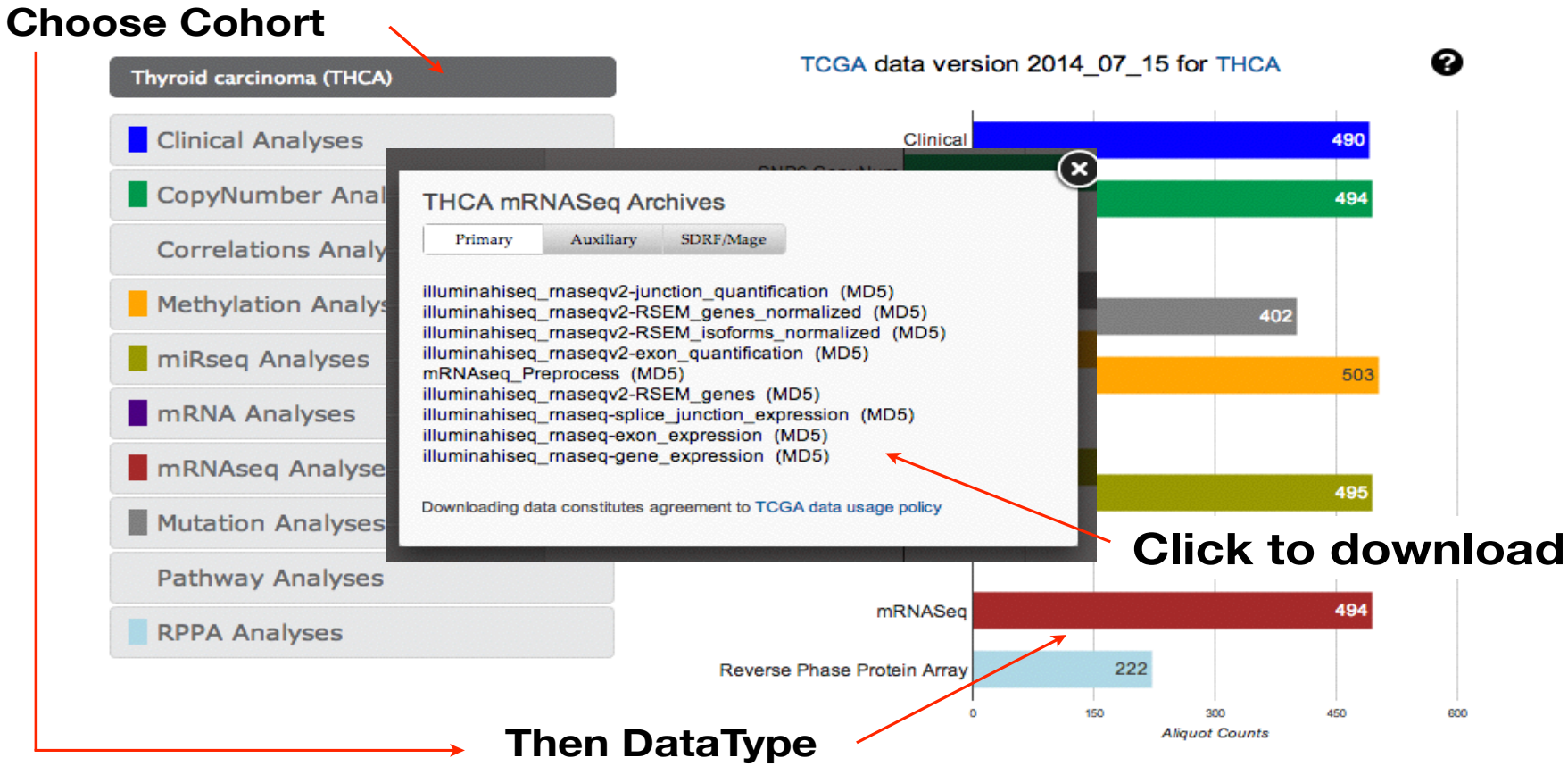
1. Access analysis reports

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

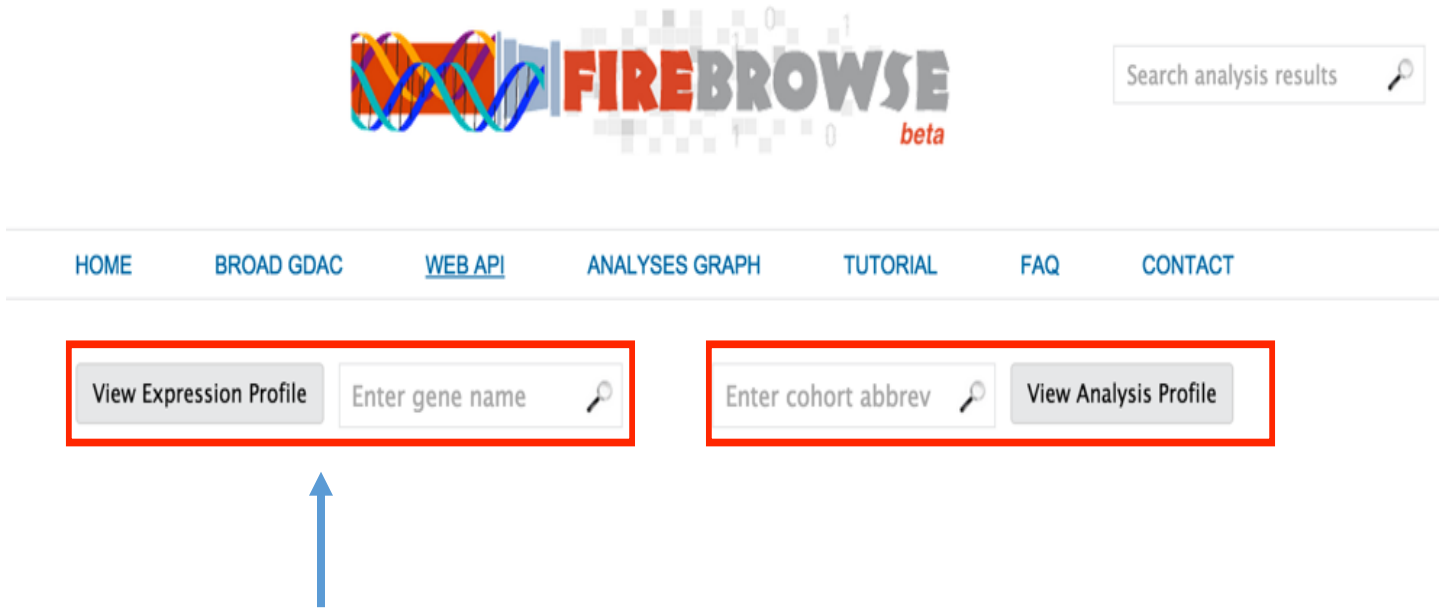


2. Access version stamped datasets

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

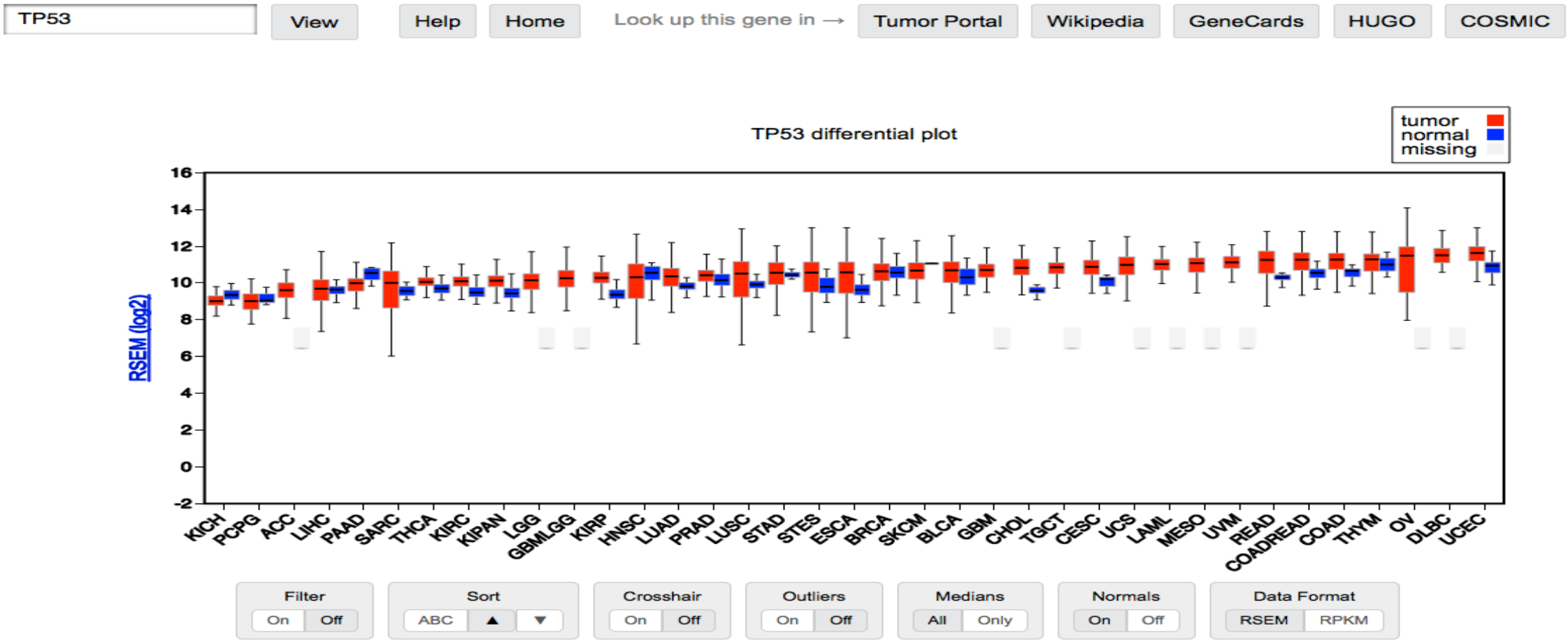


3. Visualize expression profile of one gene across cancer cohorts



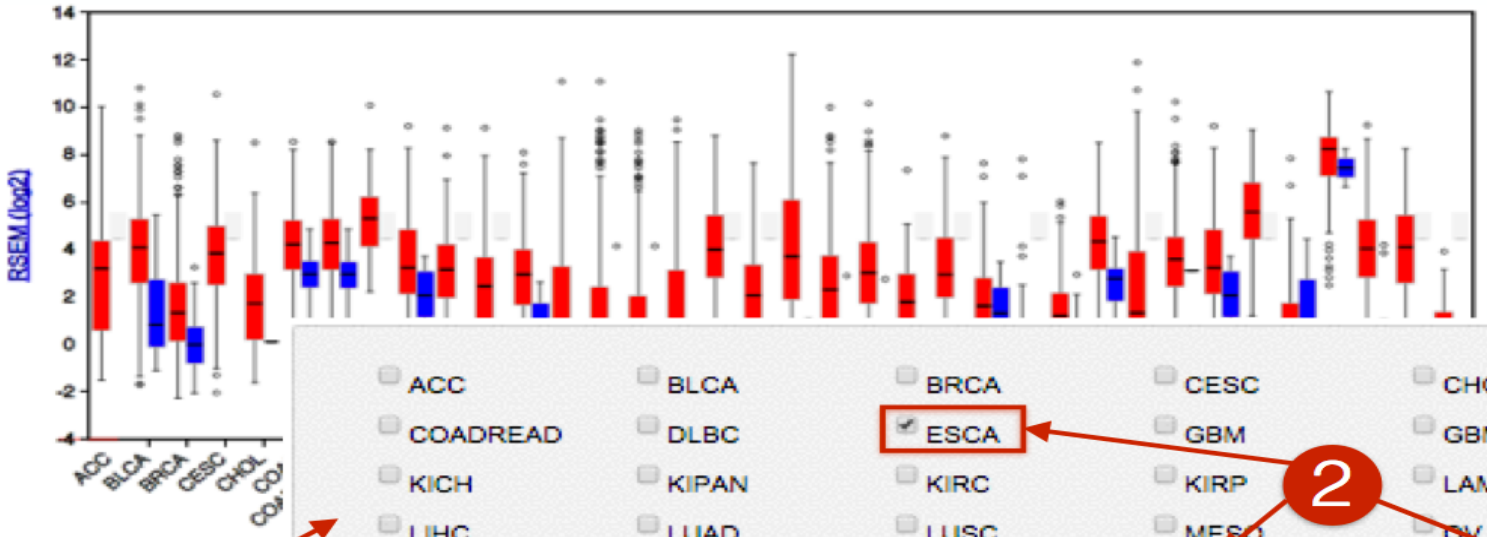
View expression profile of one gene across different cancer cohorts

Visualize boxplots of mRNA expression profiles for a selected gene across cancer cohorts

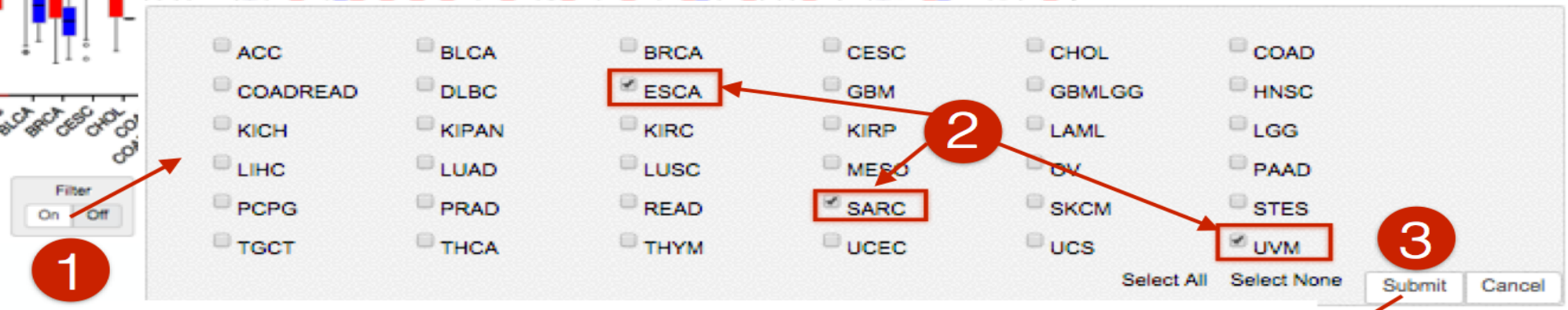


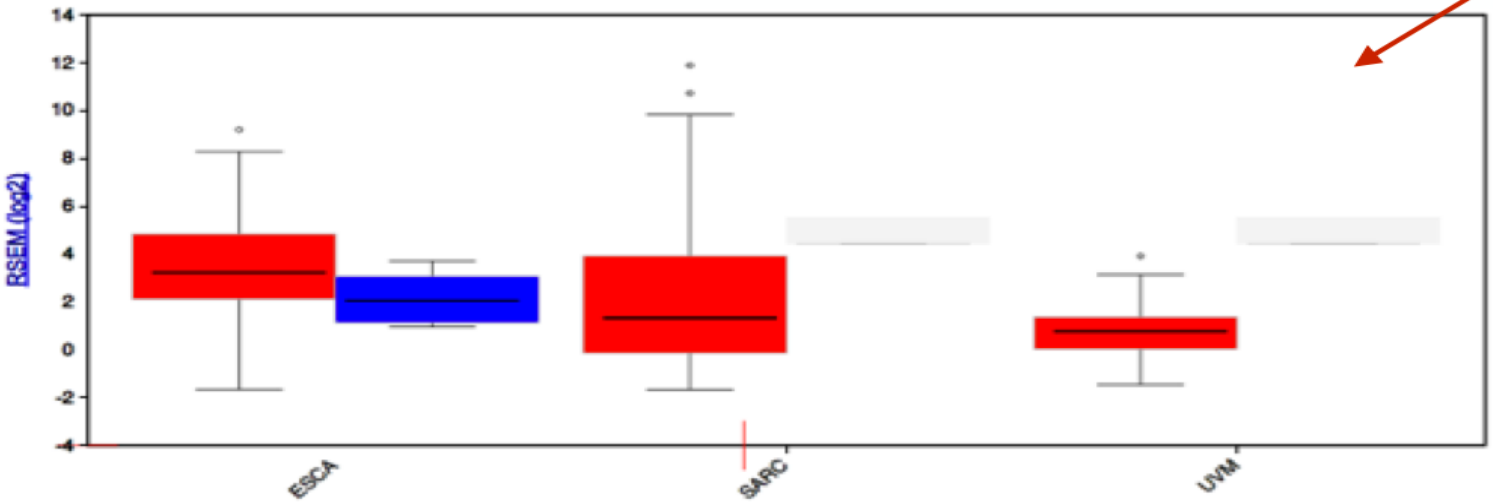
Built on top of the FireBrowse API, lets one quickly inspect mRNASeq expression levels for a selected gene, across all cohorts.

TERT differential plot



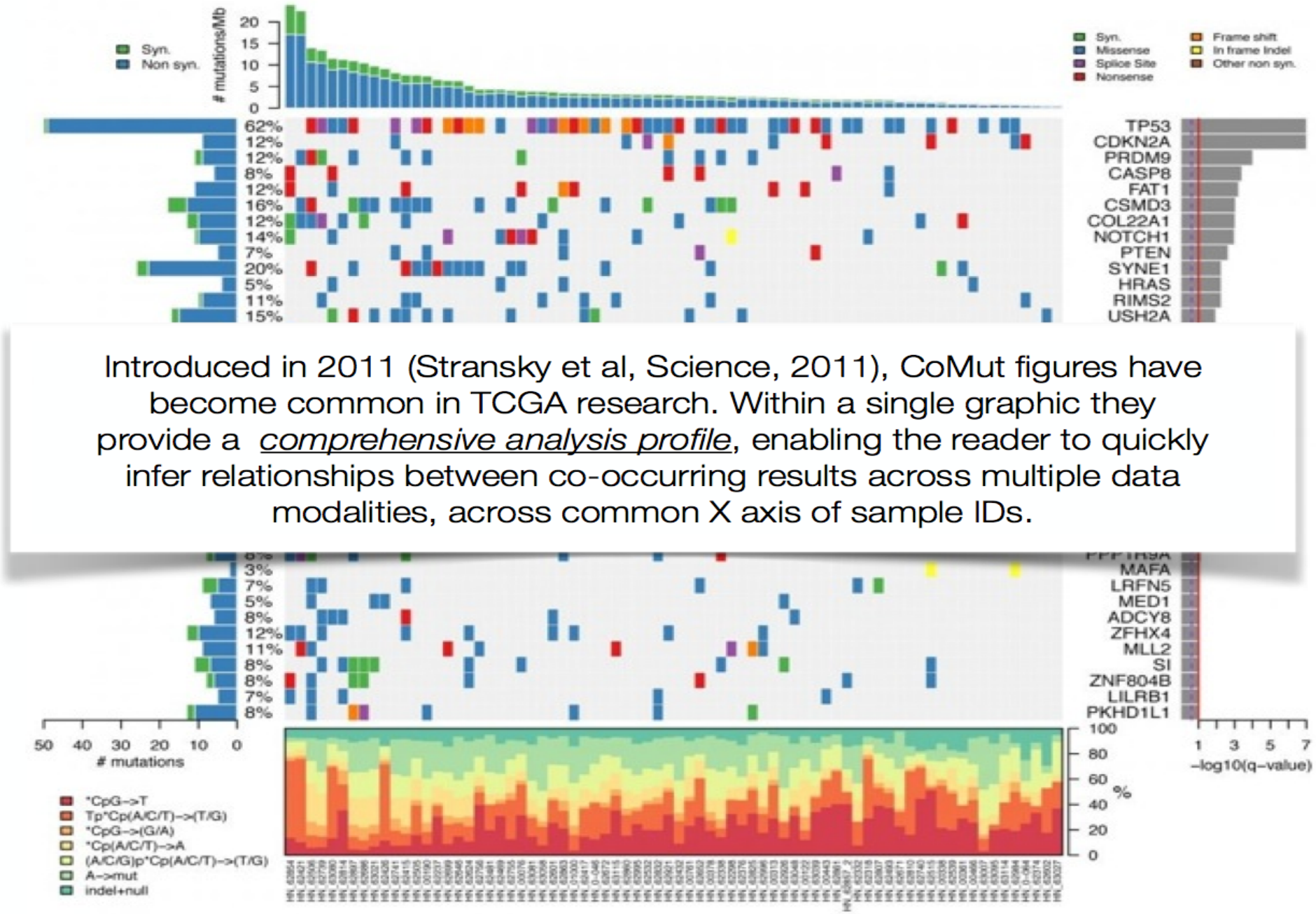
View expression levels across all cohorts, or arbitrary subsets.



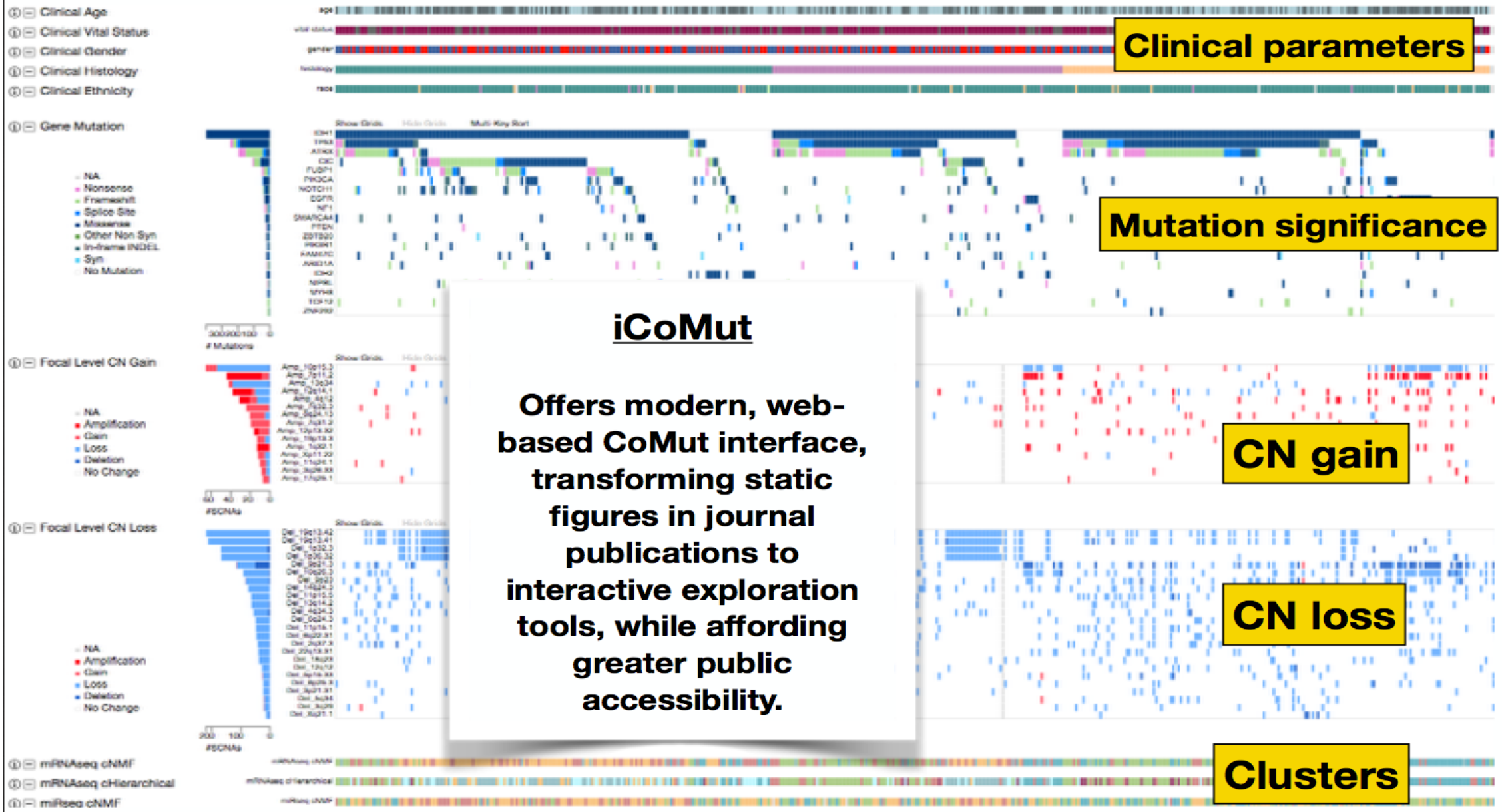


4. Visualize integrative multi-panel analysis profiles for a chosen cancer cohort

CoMut: mutation co-occurence plots

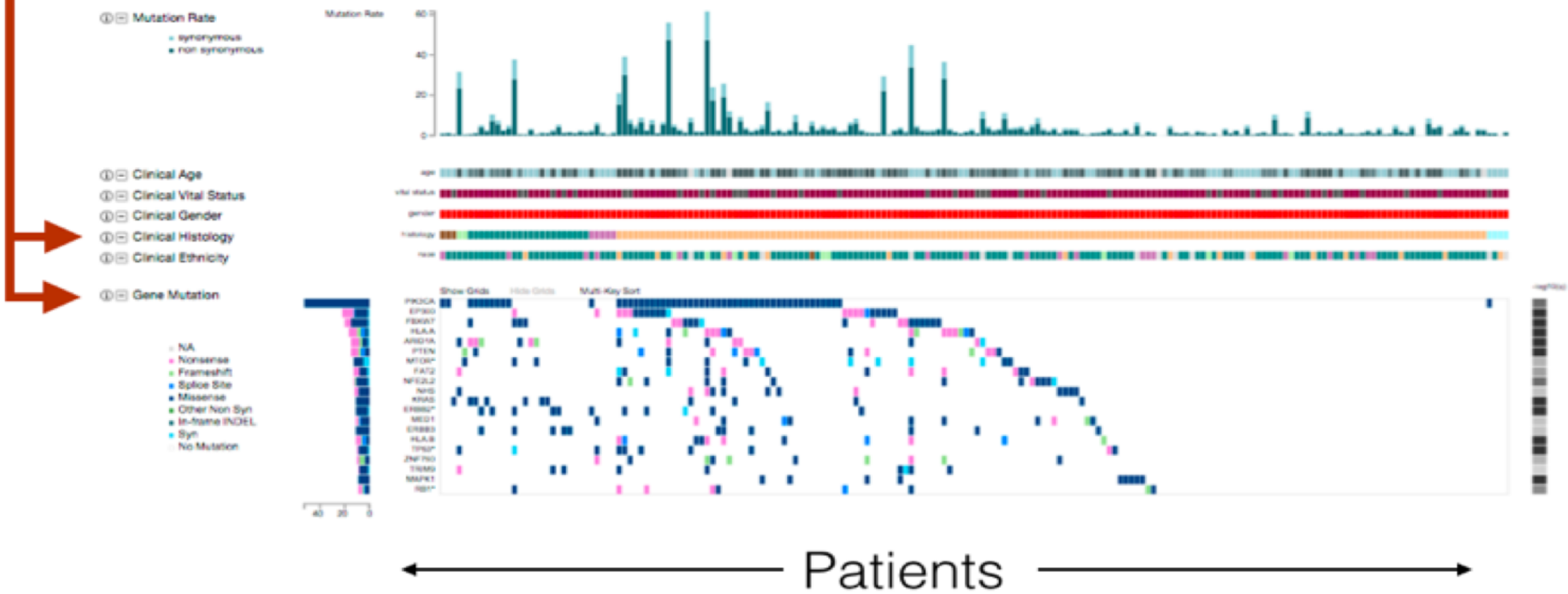


Introduced in 2011 (Stransky et al, Science, 2011), CoMut figures have become common in TCGA research. Within a single graphic they provide a *comprehensive analysis profile*, enabling the reader to quickly infer relationships between co-occurring results across multiple data modalities, across common X axis of sample IDs.



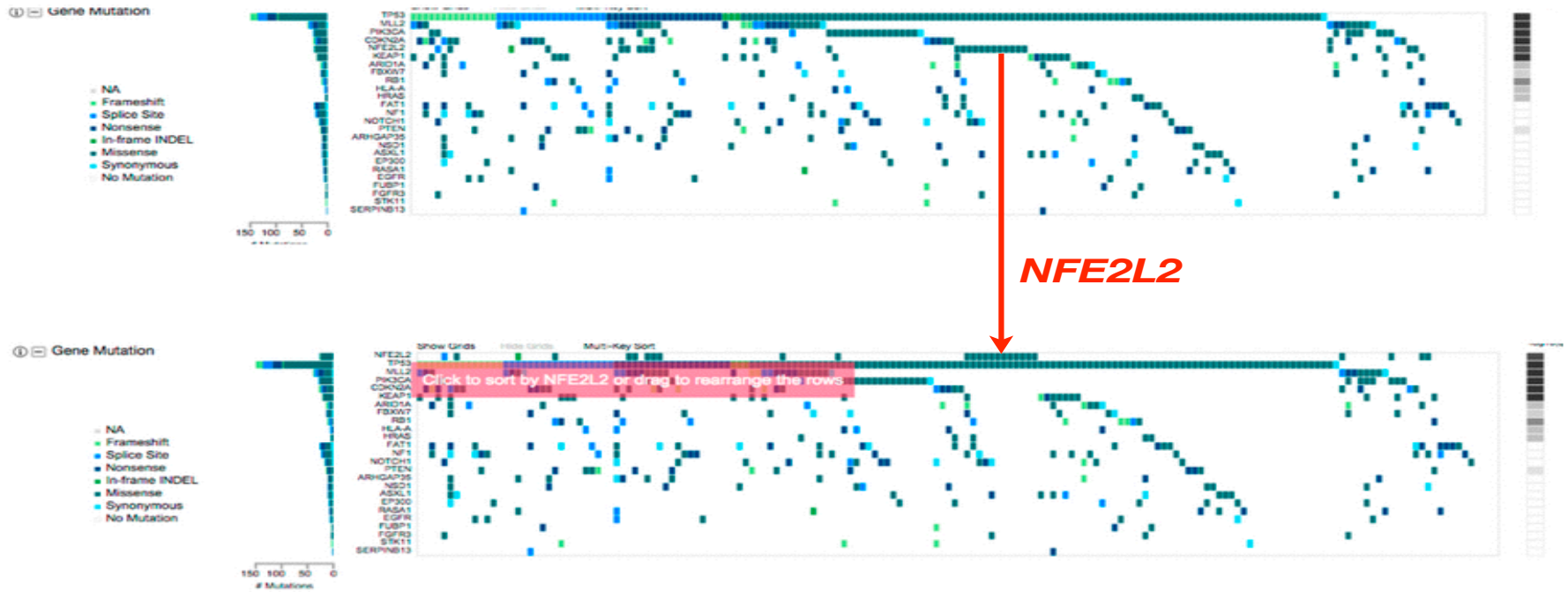
firebrowse.org/iCoMut/?cohort=LGG

By default, patients are sorted by histology and gene mutation



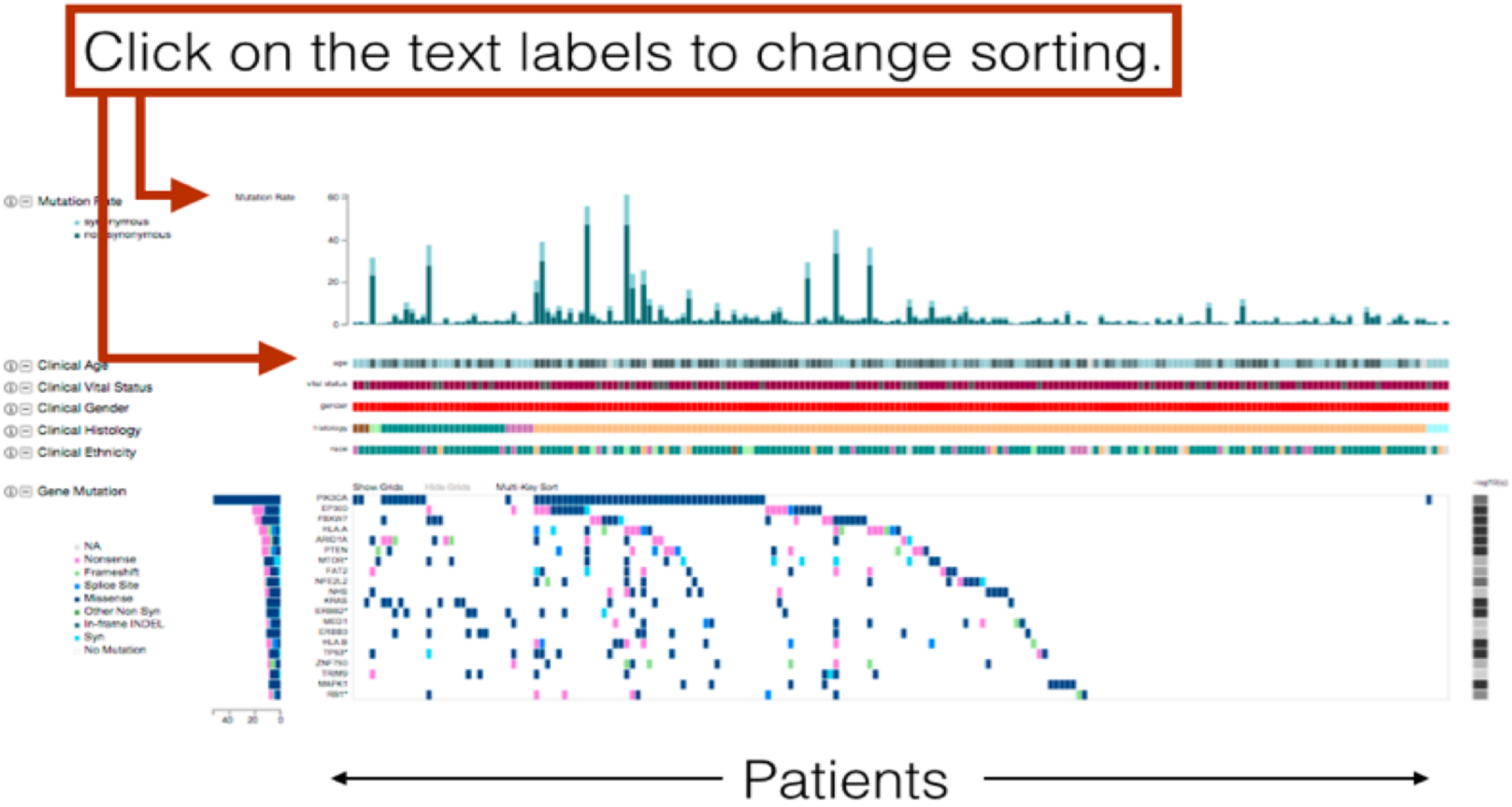
iCoMut

Drag and drop the row names to rearrange the row order

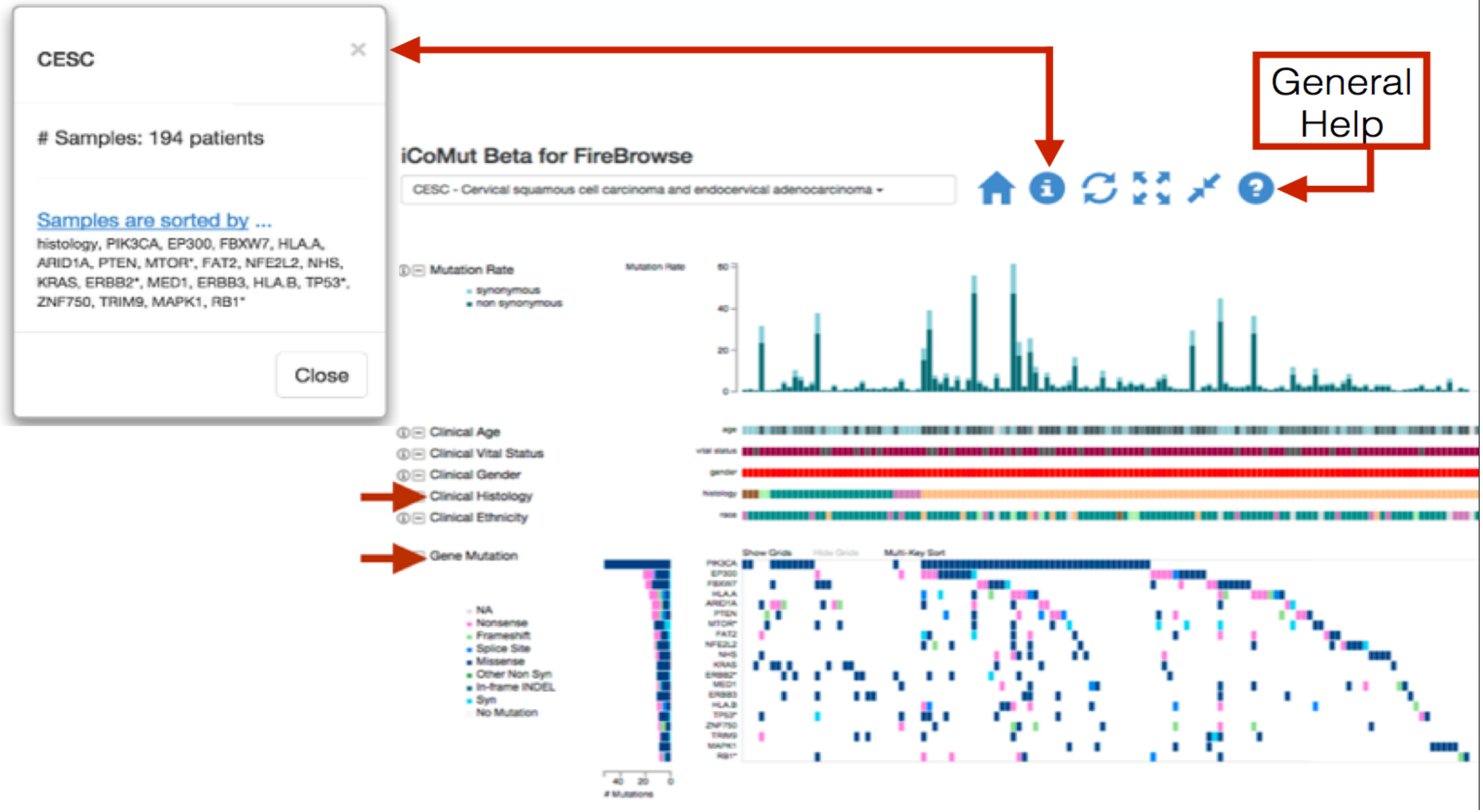


and many more graphical controls …

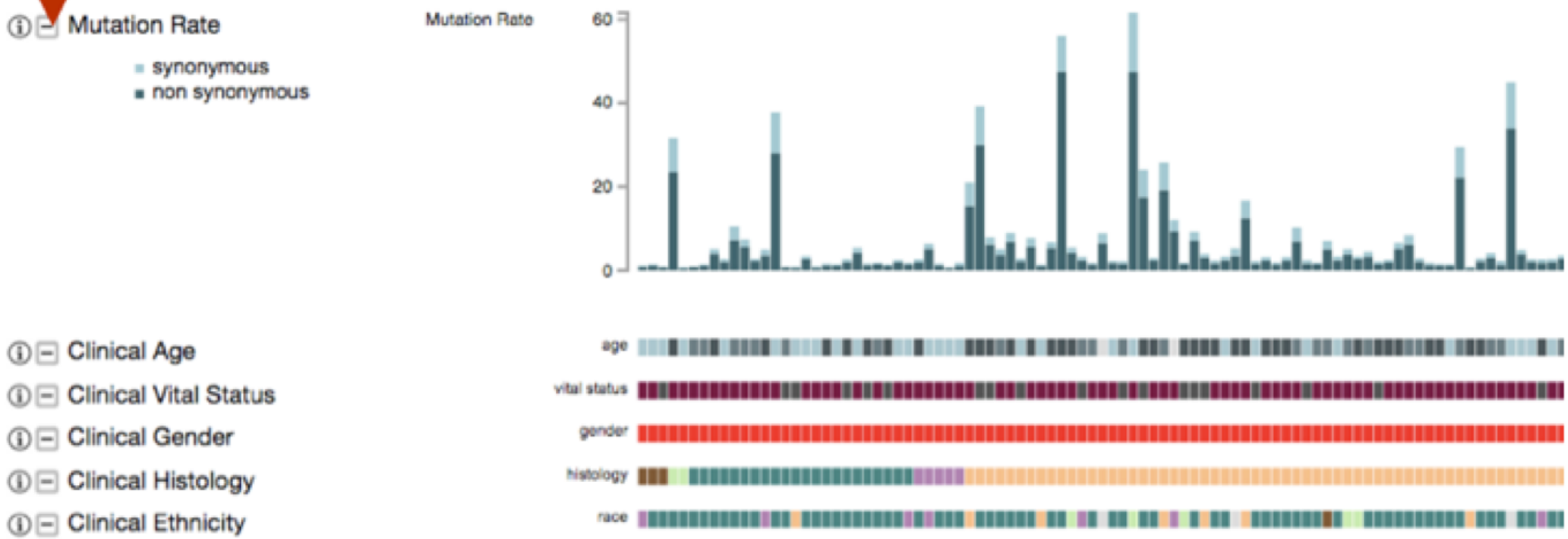
iCoMut takes researchers beyond staring at static figures in journals, wondering what the pixels mean, and how they'll reproduce —allowing them to interactively view, sort and reorder samples & results as they see fit


Click on the text labels to change sorting.
Mutation Rate
Clinical Age
Clinical Vital Status
Clinical Gender
Clinical Histology
Clinical Ethnicity
Gene Mutation
NA
Nonsense
Frameshift
Splice Site
Missense
Other Non Syn
In-frame INDEL
Syn
No Mutation
Show Grids
Hide Grids
Multi-Key Sort
Patients

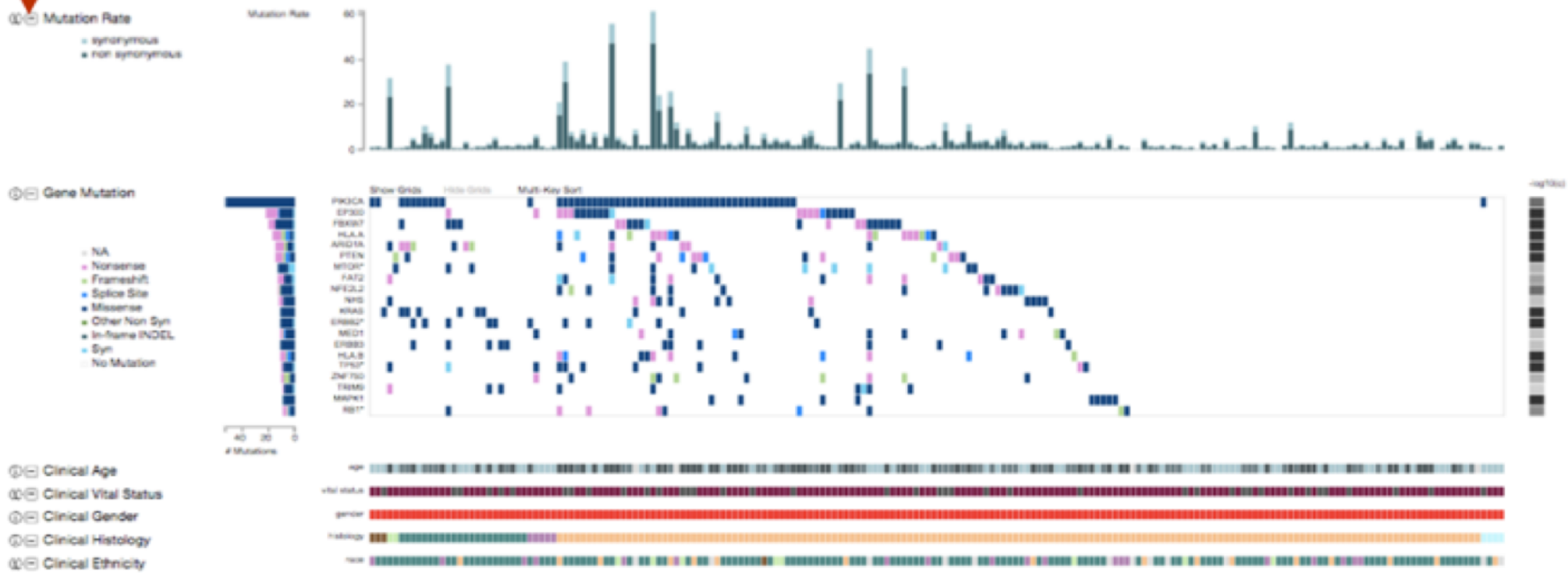
The sort status of samples is reported in the info box



Click on ⊟ to collapse a panel



Drag and drop the − or + icon to rearrange the panels



Rearranged panels

iCoMut Beta for FireBrowse

ACC - Adrenocortical carcinoma

- 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
- Methylation cNMF
- RPPA cNMF Clusters
- RPPA cHierarchical

Collapse all panels

Expand them again

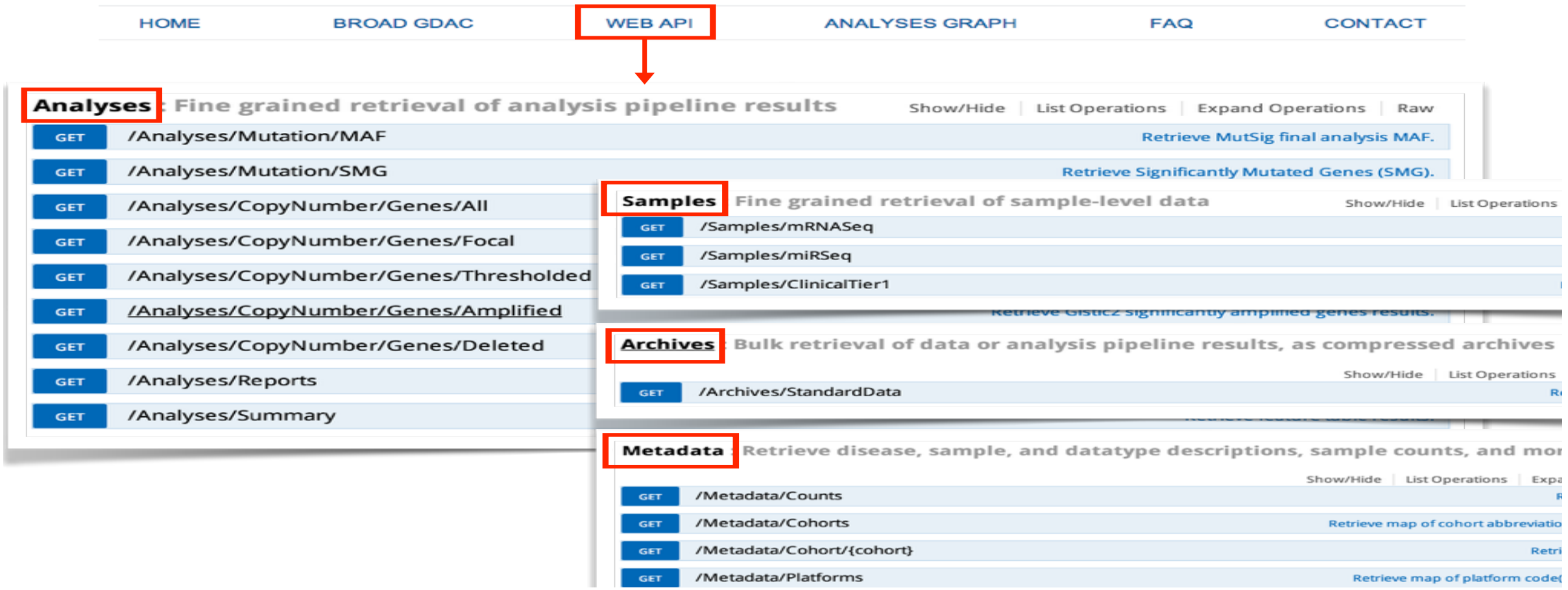
Cross hair mode

More features described in online help

http://firebrowse.org/iCoMut/#icomutHelp

5. Construct your own interactive queries

API-Powered : 25+ RESTful apis in 4 categories



Interactive Docs

learn APIs and explore data
by playing in real time
instead of cut/paste from static HTML or PDF

automatically generated & updated
as API and database evolve

GET /Samples/mRNASeq

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

choices clearly enumerated

Perform Query Hide Response

Proper RESTful call is ASSEMBLED FOR YOU

Request URL

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


```
{
  "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
},
```


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

JSON for computers/programmers

TSV, CSV for scientists, algorithms

Even Easier in Python, R, and UNIX

fbget

- Low-level Python bindings: 1-1 with RESTful api
- Higher-level interface, for easy/common bioinformatics
- UNIX command line interface, too
- Automatically generated, easily synched with RESTful API
- Flexible, copiously documented and tested
- BSD-style open source license

FireBrowseR : bindings for R

https://github.com/mariodeng/FirebrowseR

? But how do I re-analyze large scale TCGA datasets with my custom/updated analytical pipeline without having to download terabytes of data on my local computer ?

Three Cancer Genomics Cloud Pilots

Broad Institute	• PI: Gad Getz • Google Cloud • **Firehose in the cloud** • http://firecloud.org
Institute for Systems Biology	• PI: Ilya Shmulevich • Google Cloud • **Interactive visualization and analysis** • http://cgc.systemsbiology.net/
Seven Bridges Genomics	• PI: Deniz Kural • Amazon Web Services • **> 30 public pipelines** • http://www.cancergenomicscloud.org

Cloud Pilots: Coming your way!

Broad

- Version 1.0 – 1/20/2016
- Version 1.1 – 4/2/2016

ISB

- Pre-release – 11/15/2015 (open-access data)
- Version 1.0 – 12/20/2016
- Version 2.0 – 3/20/2016

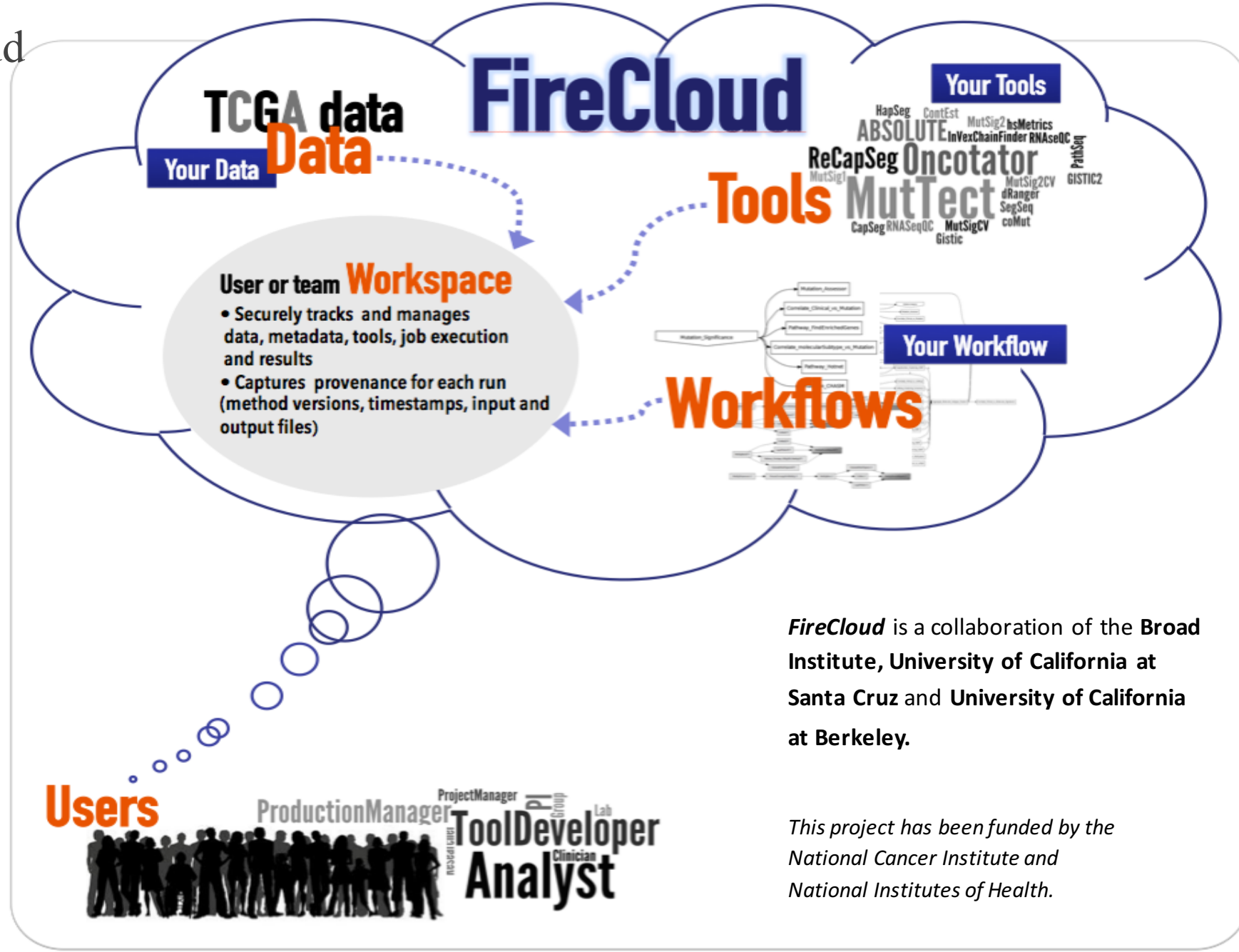
SBG

- Early access – 11/15/2015
- Version 1 – 12/28/2015
- Version 2 – 3/28/2016

Firehose re-born in the cloud

FireCloud is modeled on **Firehose**, the cancer genome analysis platform built by the Getz lab at the Broad Institute, which supports both small groups and major projects (e.g. TCGA, GTEx). ***FireCloud*** significantly expands on **Firehose's** capabilities. **Firehose** is used by both production managers for large-scale analysis and analysts for interactive analysis, curation and manual review of data for publication.

Free trial workspaces are available for all new users!



FireCloud is a collaboration of the **Broad Institute, University of California at Santa Cruz** and **University of California at Berkeley.**

This project has been funded by the National Cancer Institute and National Institutes of Health.

What is FireCloud?

A workspace environment

- holds data, tools workflows, results
- has robust security
- provides access control for users and groups
- workspaces can be shared or cloned - including tools, data, workflows and results
- provenance is captured for every run (*ie: what version of what method was run on what data at what time)*
- holds pre-loaded analyses of TCGA data (by tumor type) and other public data

FireCloud is pre-loaded with data

- TCGA data
  - public data will be accessible to all users
  - controlled access data will require dbGAP approval
- other public datasets
  - 1000 Genomes, CCLE, GTEx
- data is co-located for efficiency
  - no need to download and store large files
- users can upload data to workspaces to co-analyze with TCGA or other data

An informatics infrastructure in the cloud

- elastic compute in the Google Cloud allows for expandable compute capacity
- promotes community-wide access and use
- public REST APIs provide scalability and service
- tools and workflows are 'containerized' for security, packaging, and tracking

FireCloud contains Tools, Pipelines, and Workflows

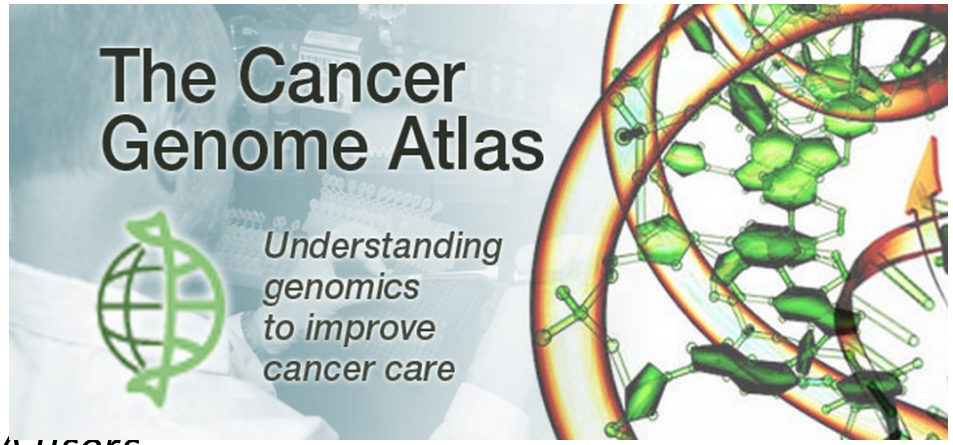
- the Method-store enables publishing & sharing tools
- workflows can be assembled with Workflow Description Language (WDL)
- best practice pipelines and other standard pipelines current state-of-the-art tools are available
- tutorials and documentation will be provided

This effort is firmly rooted in the data-sharing principles set forth by the Global Alliance for Genomics and Health (GA4GH), making it both technology-driven and mission-driven from its incipience.

What data is loaded in FireCloud?

***FireCloud** users will be able to analyze and compute on TCGA data in the cloud*

- *The FireCloud will contain co-located TCGA data (controlled access and open access data) including:*
  - de-identified clinical data
  - copy number data
  - miRNA data
  - somatic and germline mutation calls
  - DNA-seq and RNA-seq BAM files



- *Controlled access TCGA data will be available to authorized TCGA users*
  - TCGA dbGaP approval needed for controlled access data
  - data is accessed via secure authorization and authentication (ERA Commons/NIH)
- *Open access TCGA data will be available to all users*
- *Users will be able to upload their own data into the system to co-analyze with TCGA data*
- *Co-location will reduce the need for costly downloads and storage*

What are pre-loaded FireCloud workspaces?

FireCloud *will be populated with pre-loaded workspaces, which, when cloned, will allow users to replicate analysis from curated and published works*

- *A sampling of pre-loaded workspaces follows:*
  - curated TCGA Tumor Type Analysis Working Group workspaces
  - TCGA GDAC Analysis Working Group workspaces
  - TCGA PanCanAtlas analysis workspace
  - tutorial workspaces containing paired tumor and normal cell lines
  - benchmarking data to enable users to test developing tools and methods
    - synthetic bams for testing contamination
    - synthetic bams for testing mutation calling

- *Eventually, we hope to expand and add non-TCGA data - for example:*
  - CCLE analysis workspace
  - 1000G analysis workspace
  - GTEx expression data (open access data)

A sampling of pipelines:

TCGA Production

Analysis

PCAWG Pipeline

TCGA GDAC Pipeline

- Hands-on tutorial on FireBrowse

Next-Generation Clustered Heatmaps (NG-CHM)

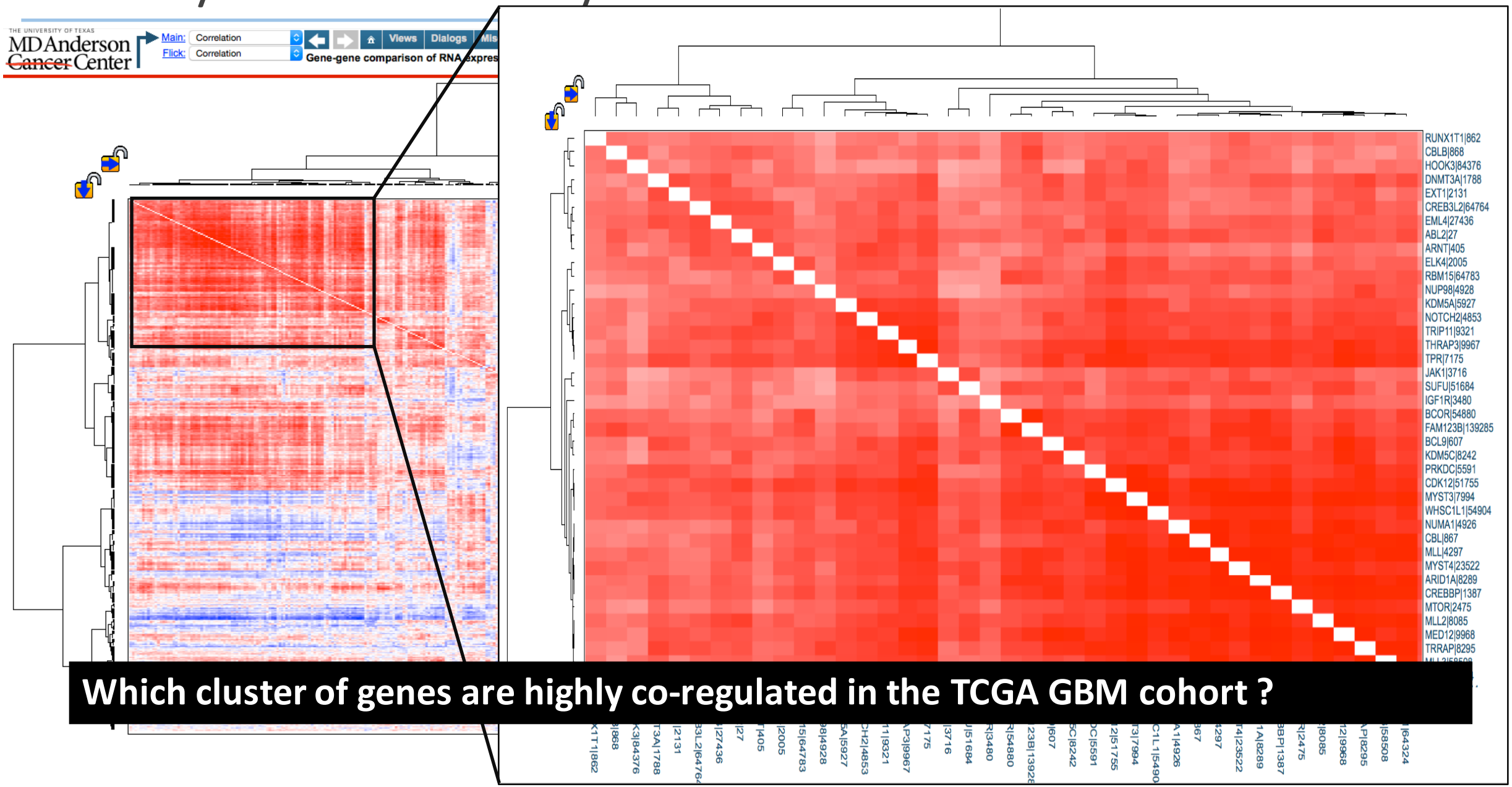
- Clustered heat maps are still the most popular way to visualize patterns in molecular profiling data, from microarrays and sequencing technologies.
- They have been included in all of the TCGA papers published in *Nature* so far on the cancers of specific tissue origin.
- Of necessity, however, they have been included in the pages of the journal as **static images**.
- ’Next-generation' clustered heat maps (NG-CHMs), use a Google Maps–like tiling technology for extreme zooming and navigation without loss of resolution.
- NGCHMs provide pathway and Gene Ontology (GO) information, chromosomal interactive ideograms, recoloring on the fly, high-resolution graphics output and linkouts to public information resources on genes, proteins, pathways and drugs.
- Perhaps most importantly, all of the metadata elements needed to reproduce them months or years later are captured and automatically saved. The result is a visually rich, dynamic environment for the exploration of the masses of data produced by TCGA.

Different ways to using NG-CHMs

- A compendium of read-only NG-CHMs showing TCGA data is available at http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/
- A publicly accessible server for creating your own test NG-CHMs is available at http://bioinformatics.mdanderson.org/testchm
- Users with advanced system administrator skills can install a local NG-CHM system instance based on docker. Installation instructions are available at NG-CHM:Docker. Once installed, users can create NG-CHMs using the built-in Builder Web Application (similar to the one on our test server).
- Advanced bioinformatics analysts can also use our high-level R library for building NG-CHMs.

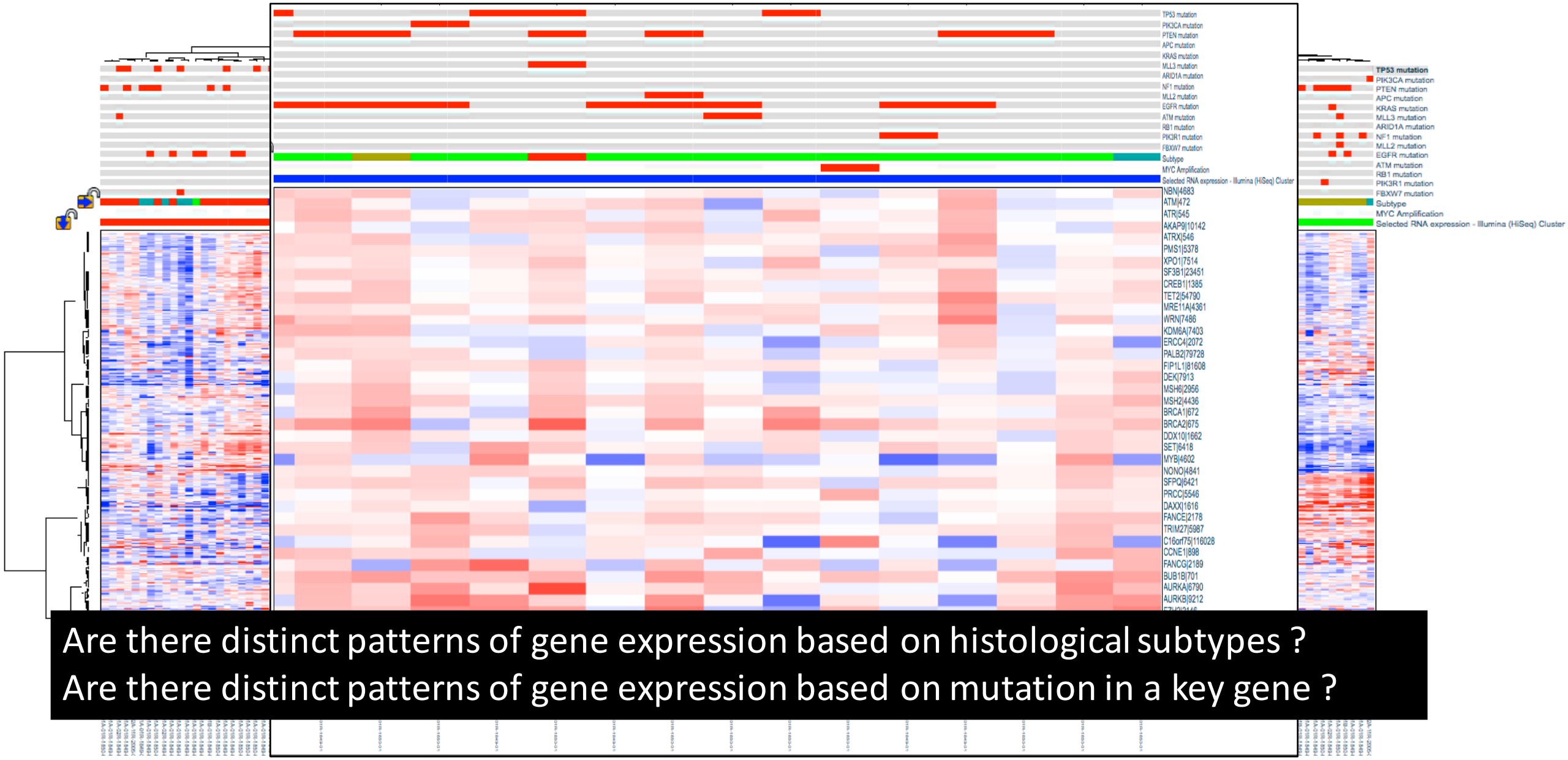
http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/

Gene/Probe vs Gene/Probe



Which cluster of genes are highly co-regulated in the TCGA GBM cohort ?

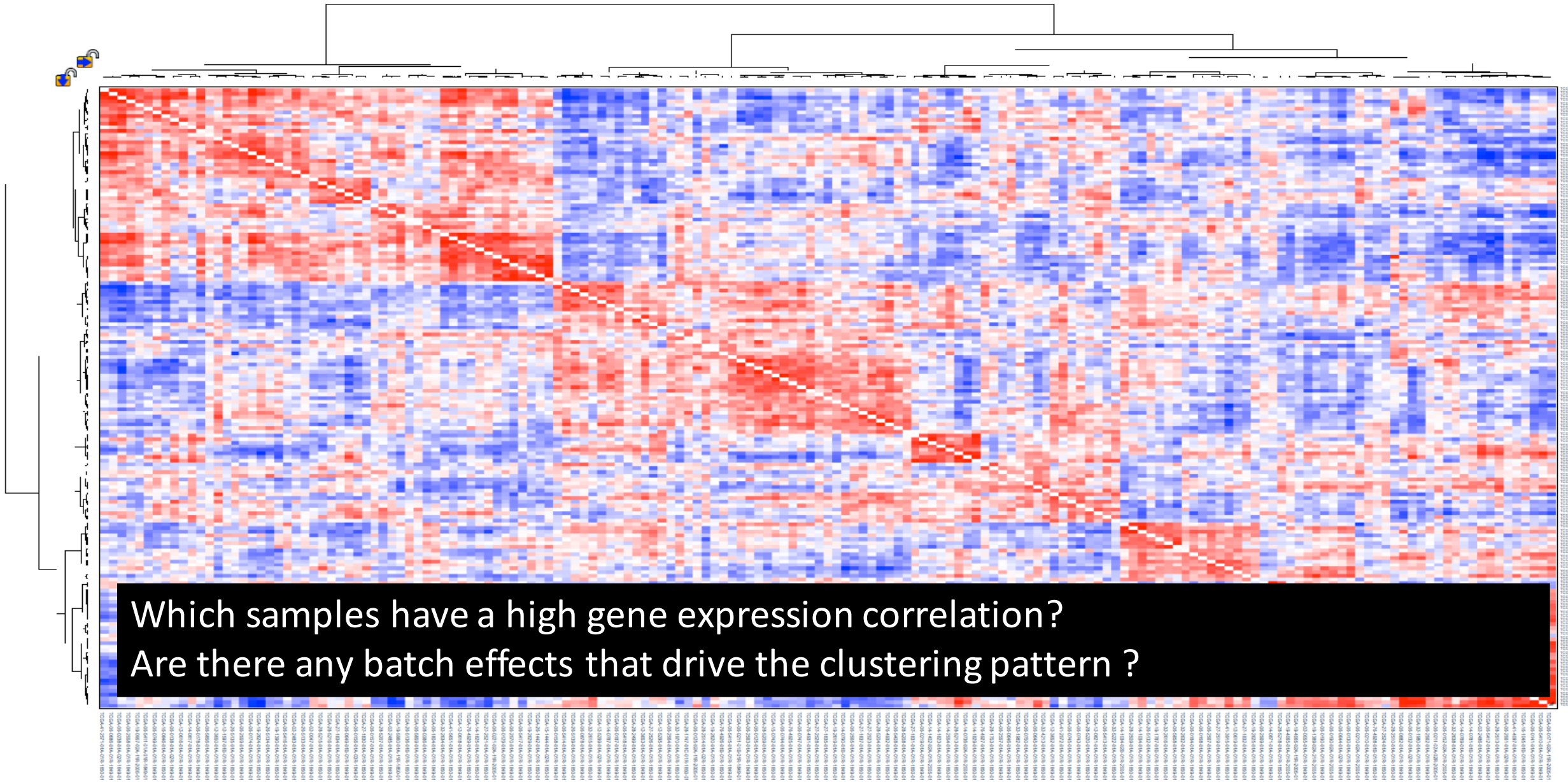
Gene/Probe vs Sample



Are there distinct patterns of gene expression based on histological subtypes ?

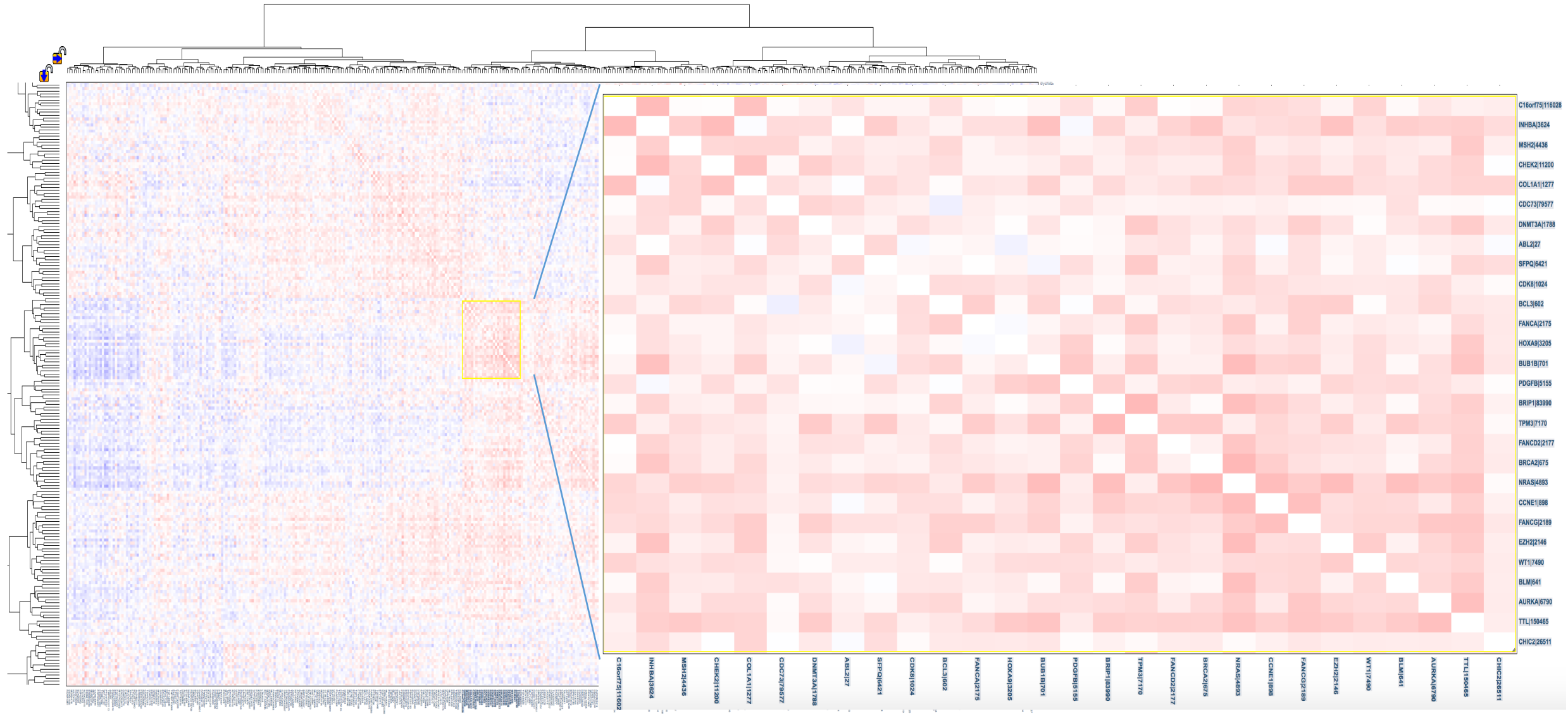
Are there distinct patterns of gene expression based on mutation in a key gene ?

Sample vs Sample

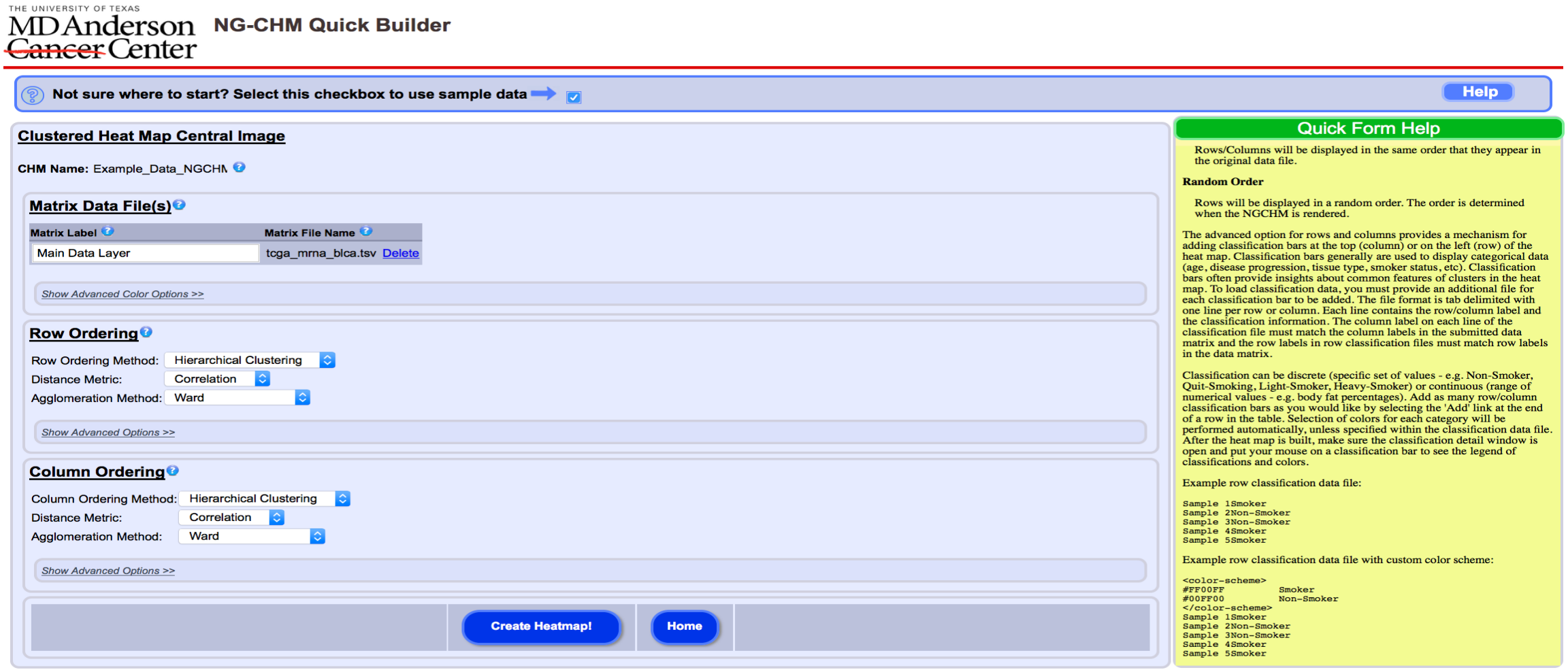


Which samples have a high gene expression correlation?
Are there any batch effects that drive the clustering pattern ?

1 x 1 Cancer Comparison Maps: LUAD vs LUSC



Create NG-CHMs for your own data



http://bioinformatics.mdanderson.org/testchm/

- Hands-on Exercise: NG-CHMs

More detailed step-by-step videos:

http://bioinformatics.mdanderson.org/main/Navigating_Clustered_Heatmaps

Using the NG-CHM Viewer

We have developed a series of tutorials and videos describing key features of the NG-CHM Viewer:

Topic	Video
Quick Tour (ppt)	N/A
Display Components and Interactivity Features	(video)
Panning and Zooming	(video)
Selections	(video)
Context-sensitive menus	(video)
Color schemes	(video)
Covariate Bars	(video)

Thank You !