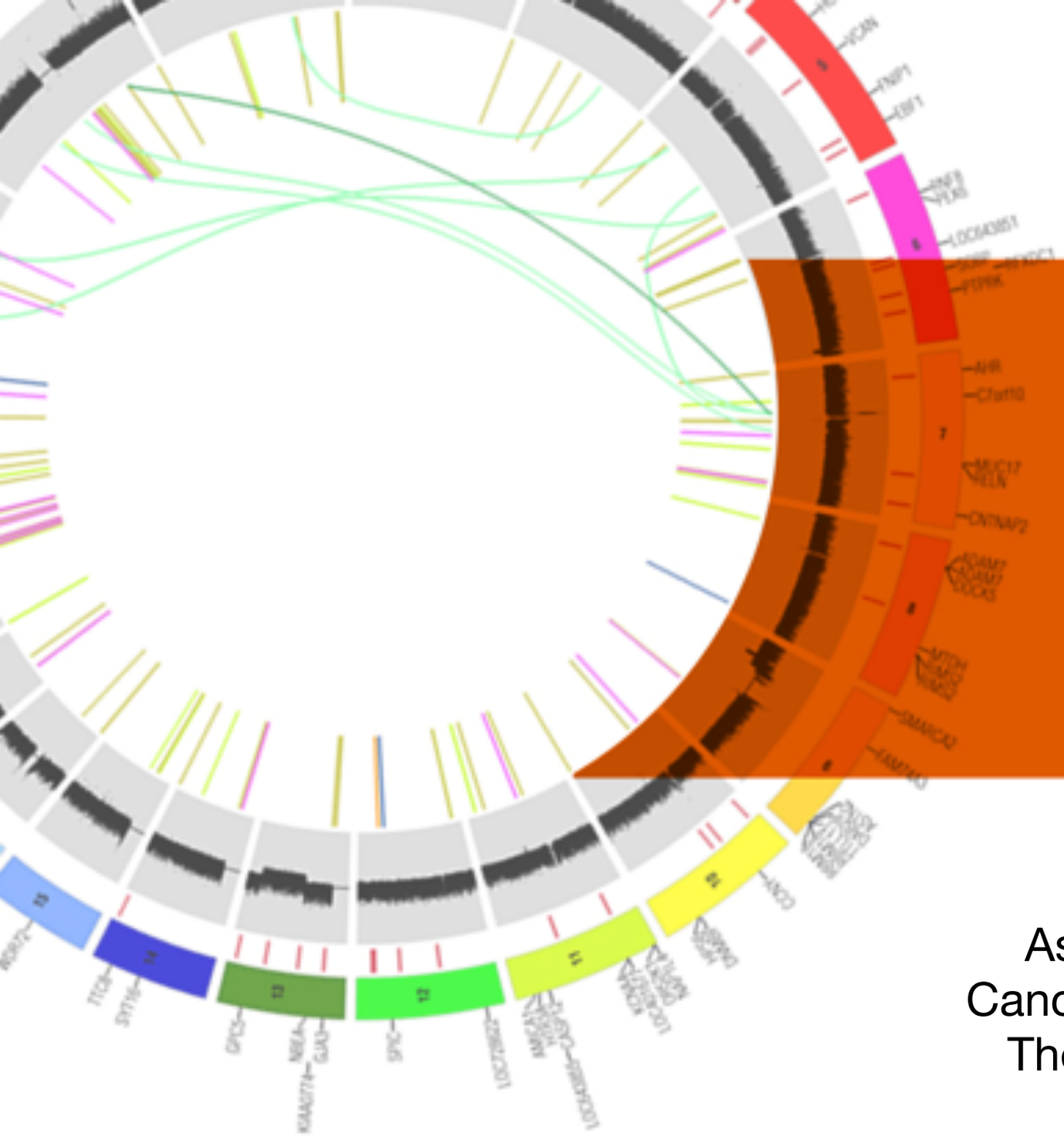


The Cancer Genome Atlas

FireBrowse: Mining Firehose of TCGA

4th TCGA Symposium
May 12, 2015
National Institutes of Health
Bethesda, MD

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

Manager
TCGA Genome Data Analysis Center

BROAD INSTITUTE

MASSACHUSETTS GENERAL HOSPITAL CANCER CENTER

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

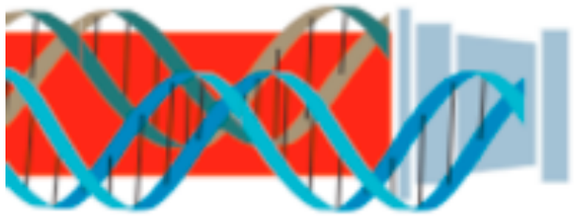
David Heiman	Hailei Zhang
Katherine Huang	Juok Cho
Kane Hadley	Jaegil Kim

The front line computational biologists
and software engineers.

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

Retrospective

(since this is our last dance)



FIREHOSE
Broad GDAC

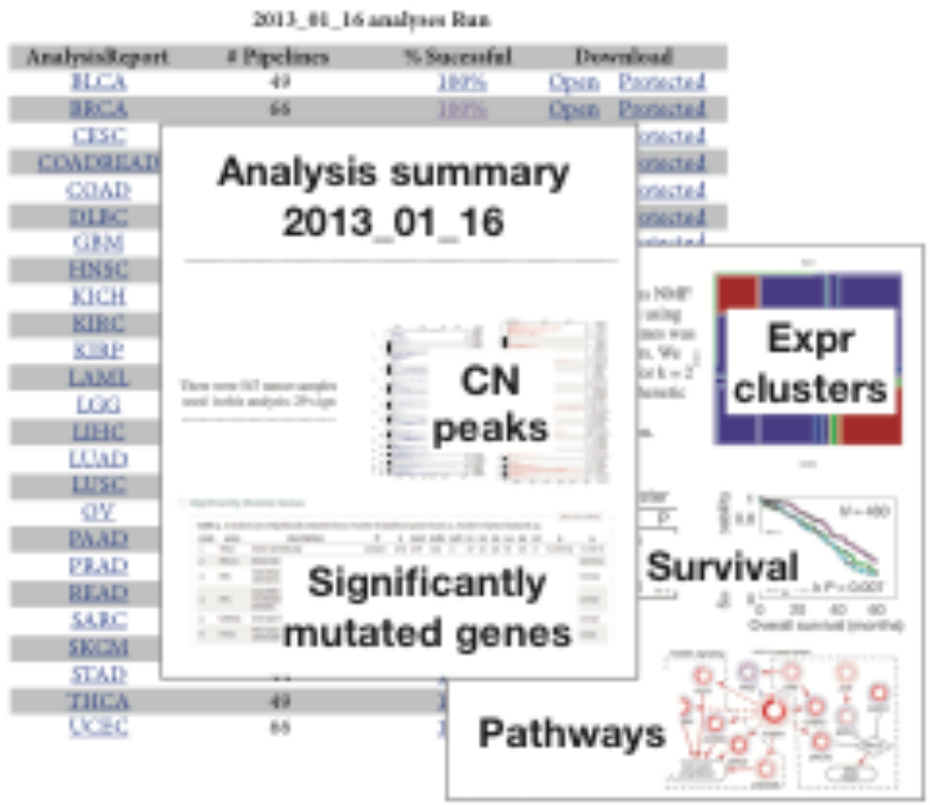
Born of the desire to systematize analyses from The Cancer Genome Atlas pilot and scale their execution to the dozens of remaining diseases to be studied, now sits atop >54 terabytes of TCGA analysis-ready data and reliably executes thousands of pipelines per month.

* Just couldn't keep doing analysis the (bad) old manual way.

Acute Need for Automation, Systematic Rigor, and Transparency

Data Factory

FIREHOSE
Broad GDAC



Significant democratizing influence of lowering entry barriers to TCGA

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

TCGA Level 1

Level 2

Level 3

Level4

FIREHOSE
Broad GDAC

publication

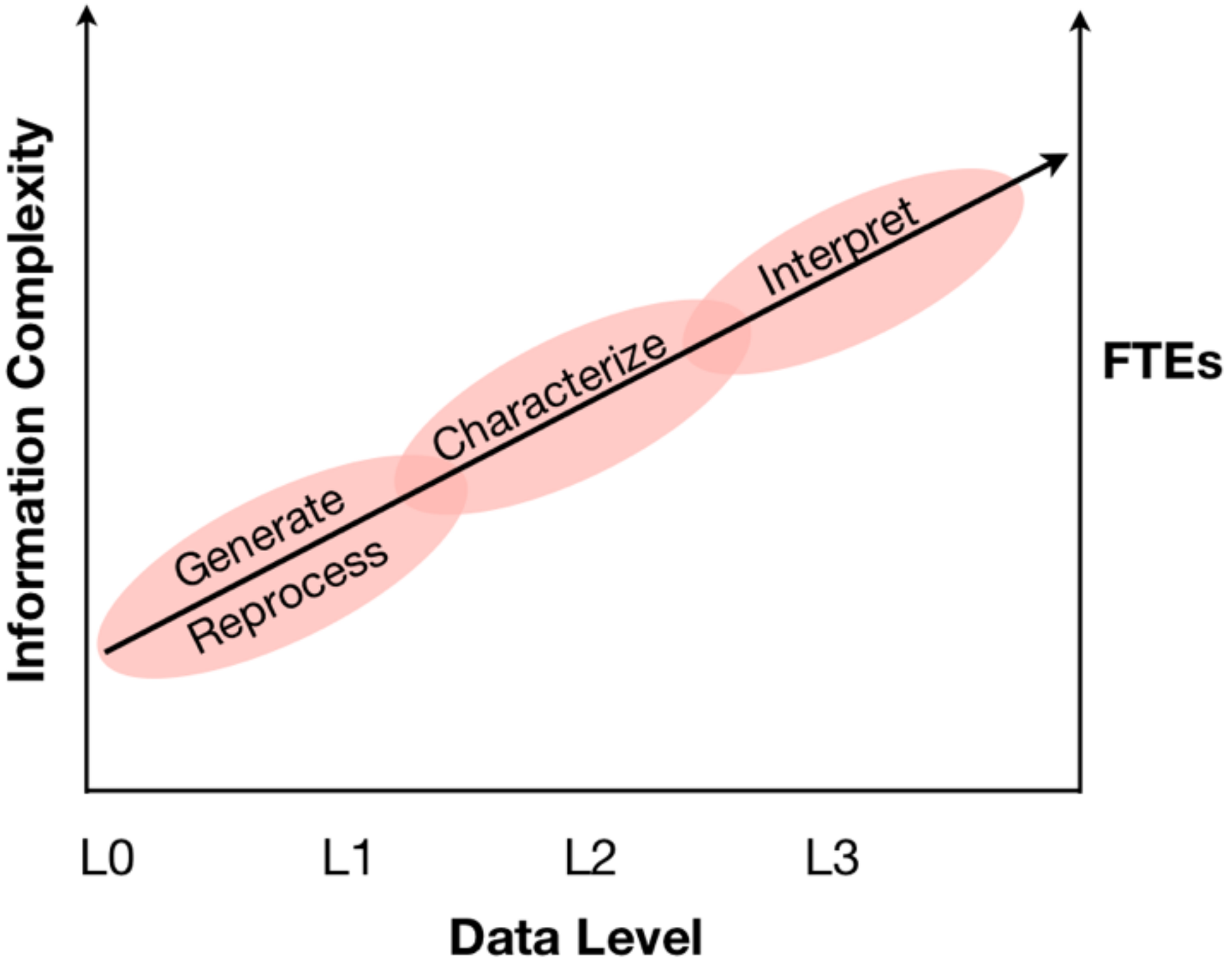
Level 5
Integrative
Analyses

Petabytes

Terabytes

Gigabytes

Megabytes



Data compression —> information density goes up —> the work gets more complex…

The Past : First Prototype Run

Parent Directory		-
2010_12_23_TumorData-annotated.png	31-Dec-2010 09:29	148K
2010_12_23_TumorDataSummary.png	31-Dec-2010 09:29	95K
Dec23_2010_Summary.pdf	31-Dec-2010 09:30	333K
run.note	31-Dec-2010 09:30	1.3K
summary.key/	31-Dec-2010 09:30	-
summary.xls	31-Dec-2010 09:30	413K
summary1.png	31-Dec-2010 09:30	112K
summary2.png	31-Dec-2010 09:30	63K

gdac.broadinstitute.org/runs/analyses__2010_12_23/

Summarized in manually crafted, 3-page PDF
.. small handful of files posted FTP style

Tumor Type	Analyses Completed	Not Completed	Percentage
OV	25	0	100%
GBM	15	10	60%
BRCA	8	17	32%
COAD	8	17	32%
LUSC	8	17	32%

- 25 tasks / workflow (many were simply preprocessors)
- Only OV cancer completed (with some elbow grease)
- 13 disease cohorts with at least 1 patient
- But even these were still very sparse
- Only OV had mutation samples
- Zero mirSeq or mRNASeq aliquots

The Present

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 Future : Spring 2016

GDAC-style workflow as early proof of concept

Scale up as far up as up goes!

You set the analysis knobs, not us!

Back on earth, GDAC Firehose produces …

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: CISTIC, 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

All downloadable with
a single command

```
linux%  firehose_get  analyses  latest
```

And because that was working well …

4 Custom runs tailored to TCGA AWGs

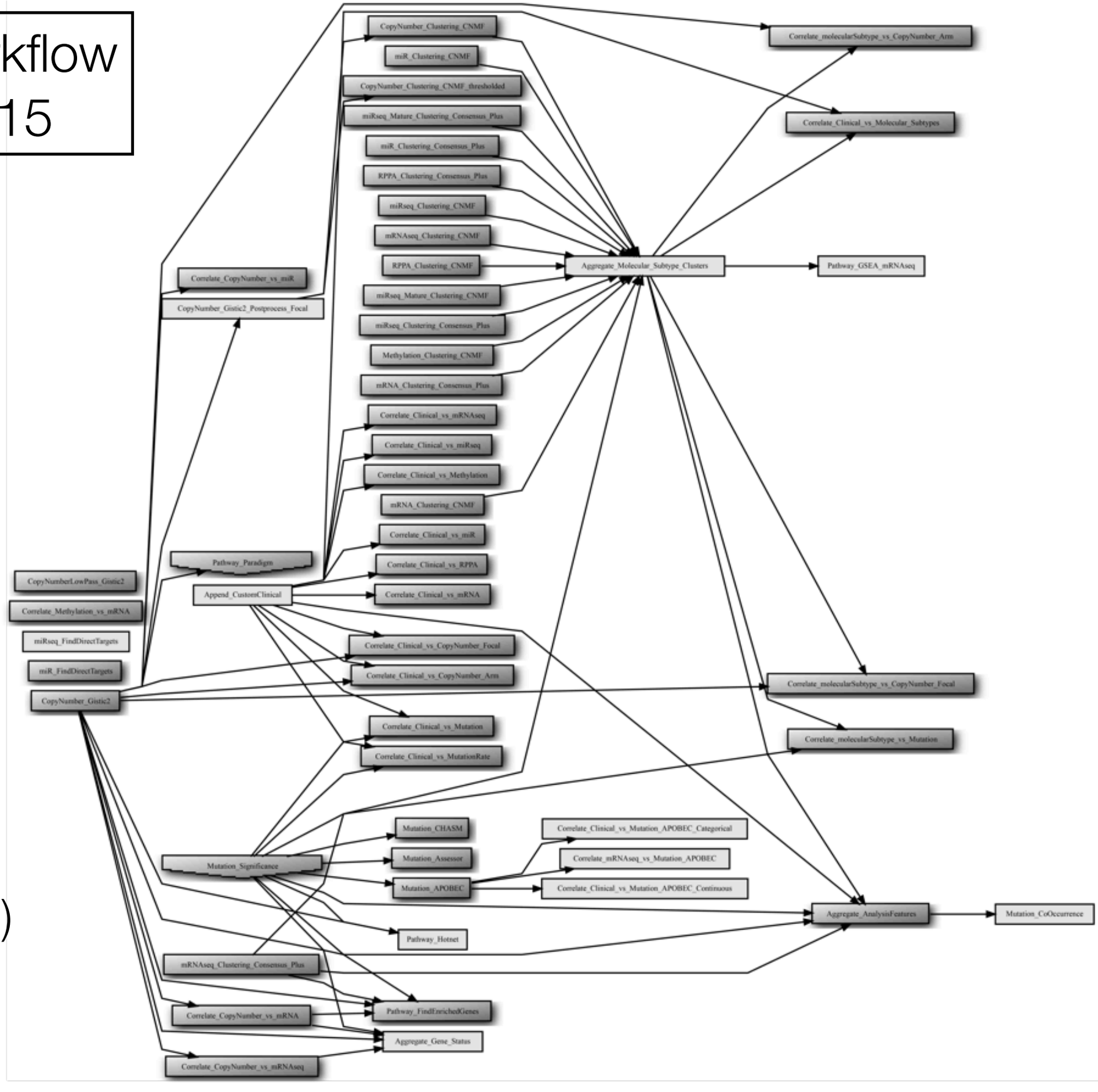
Currency:	pipelines can be run on the *latest snapshot of data* from DCC, *avoiding the time & sample lag of monthly runs*
Flexibility:	easily include AWG-curated disease subtypes, even custom analyses
Speed:	usually executed in only a few days time
Familiarity:	using same internal Firehose machinery, external-facing dashboards, Nozzle, `firehose_get` etc known to community

They look like younger sibling of Analysis runs
(and you can `firehose_get` them, too)

Analysis Workflow Spring 2015

~100 tasks total

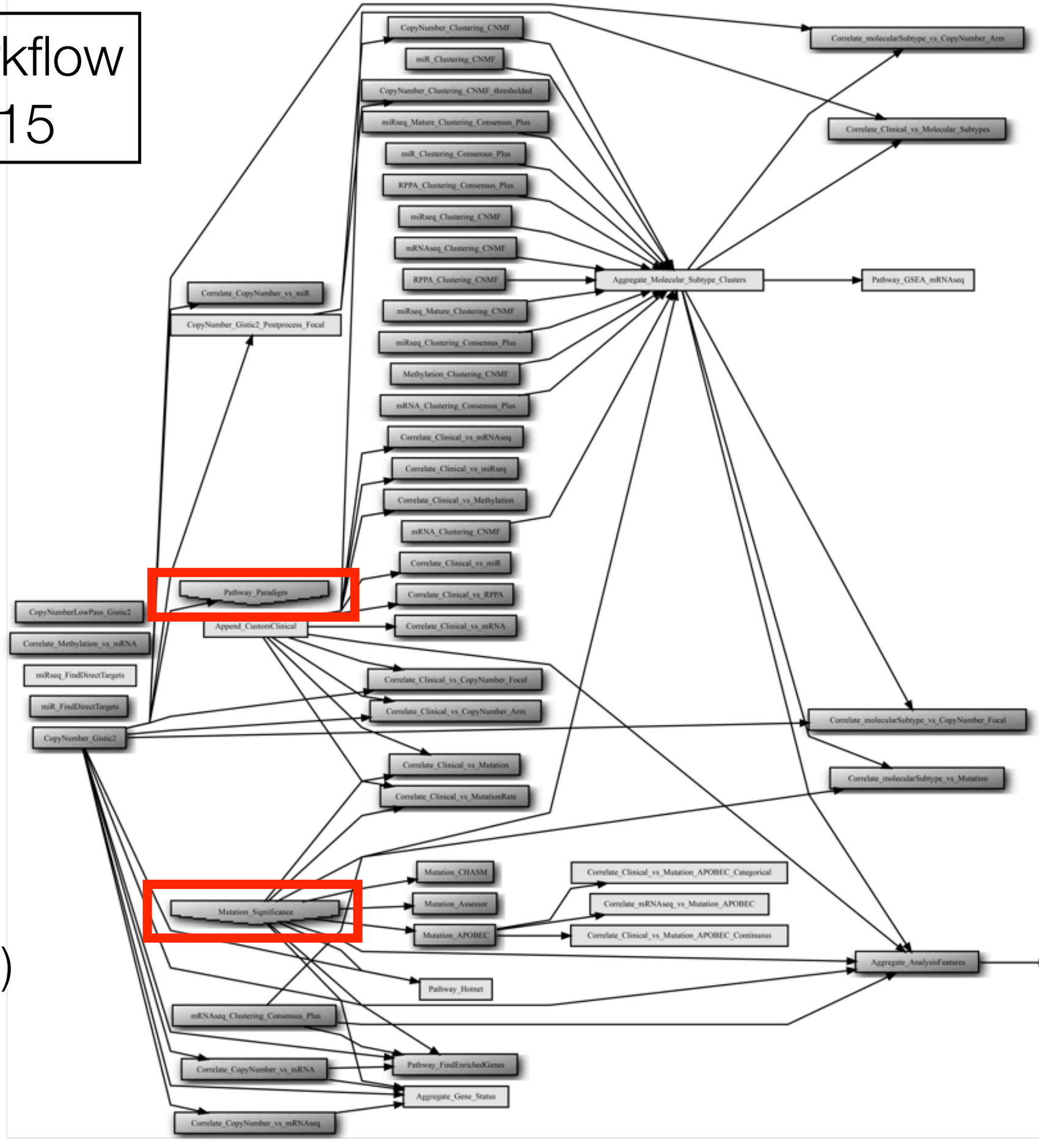
Can inject custom data (not just from DCC)



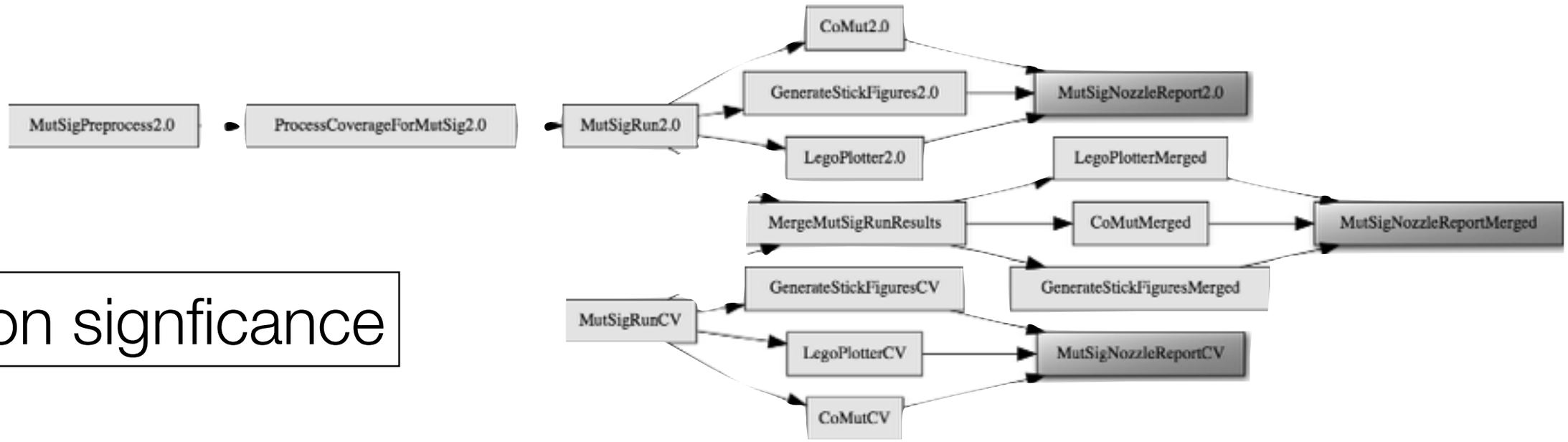
Analysis Workflow Spring 2015

~100 tasks total

Can inject custom data (not just from DCC)

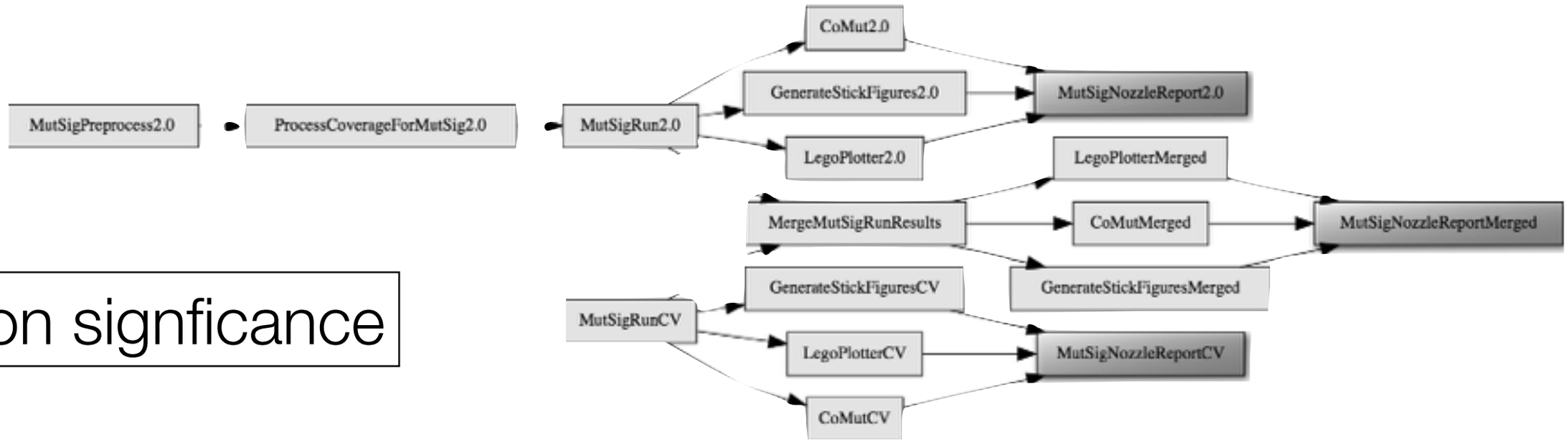


Trapezoidal nodes are subworkflows

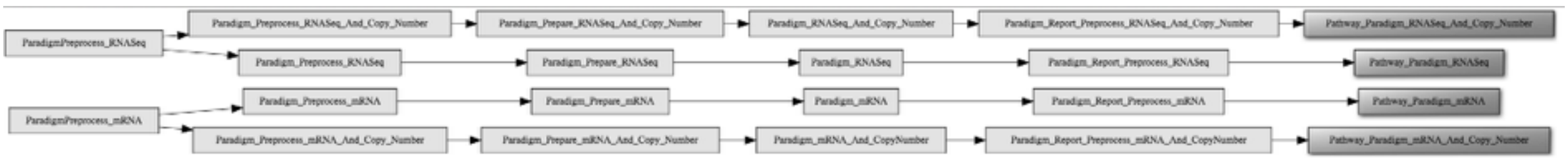


Mutation signficance

Trapezoidal nodes are subworkflows

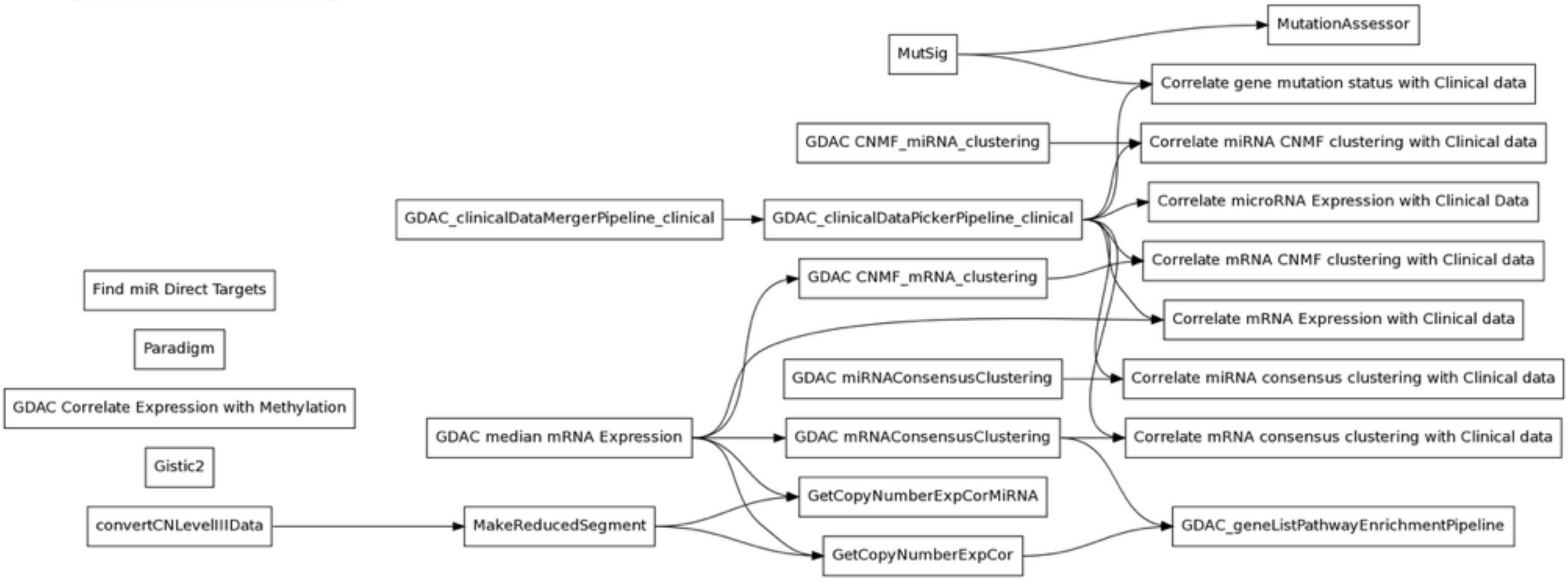


Mutation signficance



Paradigm Pathway Analysis

Contrast: entire workflow of first run



What Analyses?

- *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*

Crown Jewels
GISTIC & MutSig
(CopyNumber & Mutation significance)

  - **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,

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 (arguably most important)
vs clusters,
datatype vs. datatype
even custom data ...

- *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.

Automated, High-Throughput Clinical Miner

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

The results include survival curves for every TCGA disease (where applicable), and 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)

Fabricated aggregate cohorts

COADREAD = colon + rectal
Glioma = glioblastoma + lower grade glioma
KIPAN = kidney renal clear, papillary, chromophobe
STES = stomach + esophageal

As analysis-ready data packages …
… and comprehensively analysed results.

Plus dozens of subohorts defined as AWG subtypes

You won’t find either of these at DCC
(or any other public site in this streamlined a form?)

What's new in GDAC Firehose?

Raw MAFs

Latest analyses run includes post-publication mutation data, ***adding over 1500 mutation samples to our data stream.***

New Analyses

Correlate_Clinical_vs_Mutation_APOBEC_Categorical
Correlate_Clinical_vs_Mutation_APOBEC_Continuous
Correlate_mRNAseq_vs_Mutation_APOBEC
miRseq_FindDirectTargets
Mutation_CoOccurrence
Pathway_GSEA_mRNAseq
Pathway_Overlaps_MSigDB_MutSig2CV

(311 new reports, 1480 total; see release notes for details)

Ok, that’s all well & good …

But as reminded by DARPA yesterday* …

It’s just too much …

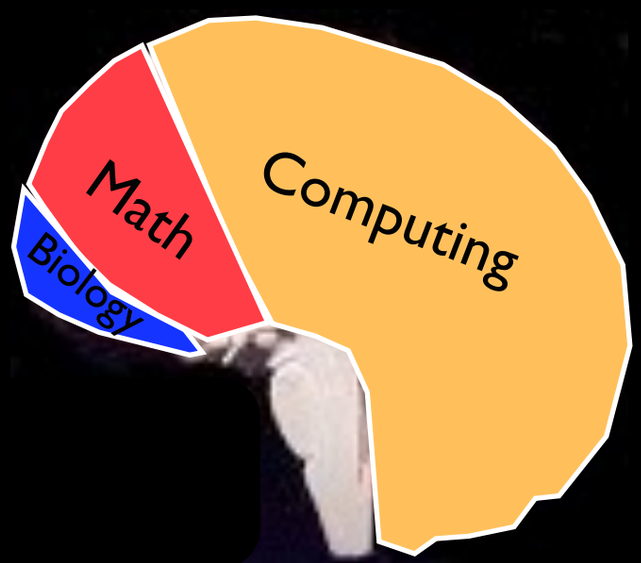
Our attempt to make things easier with Firehose needed to get even easier

* Paul Cohen: *Machines That Construct Cancer Pathways by Reading the Primary Literature*

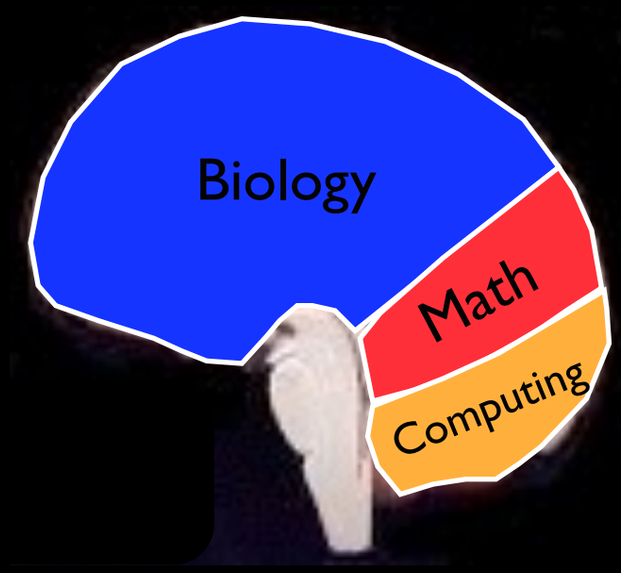
But we knew this already …

This is Your
Researcher
Brain

But we knew this already …

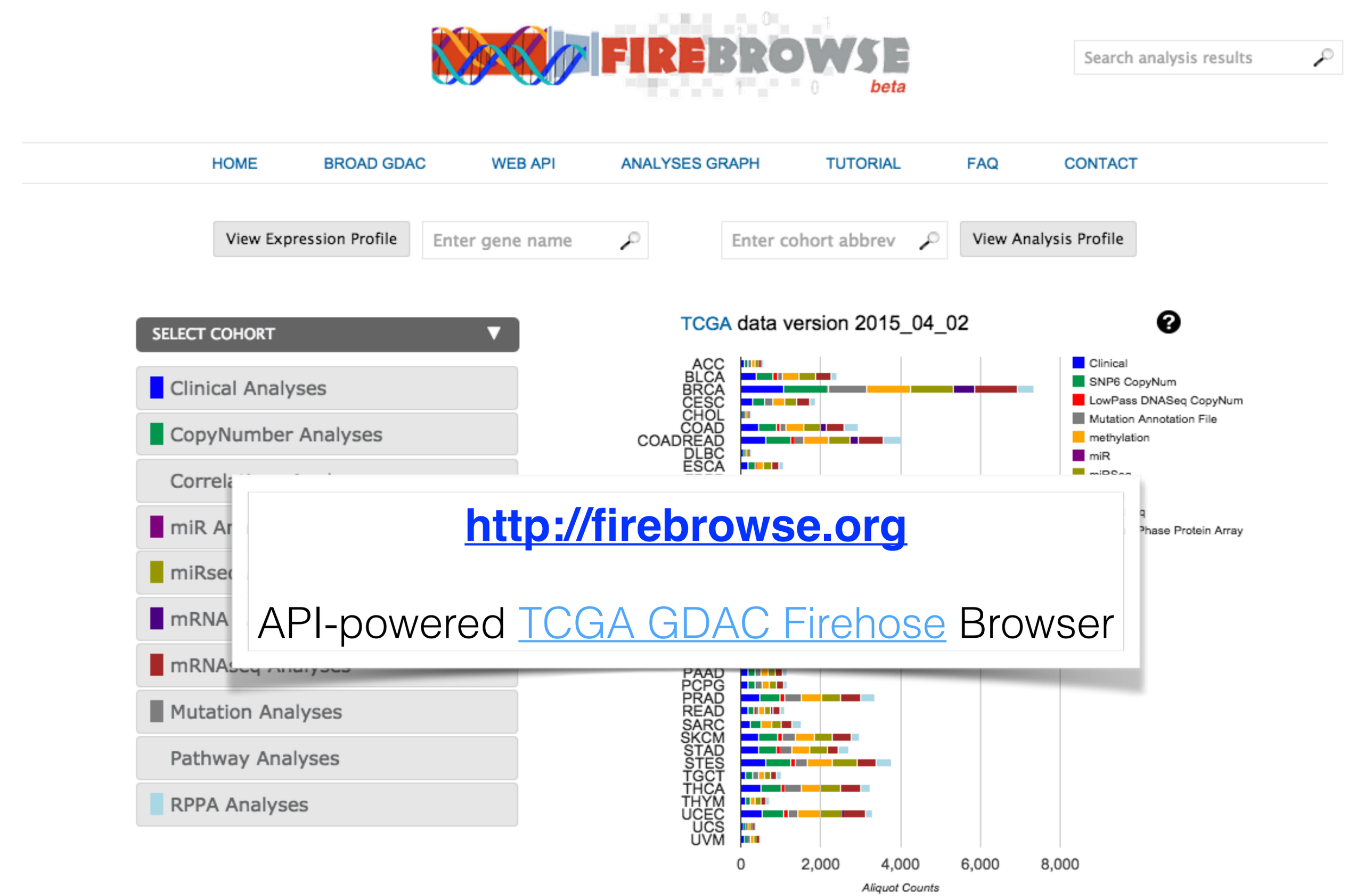


When Coding
Or Data
Exploration
Is Hard



When
Easier

Civilization advances by extending the number of important operations which we can perform without thought. **A. North Whitehead**



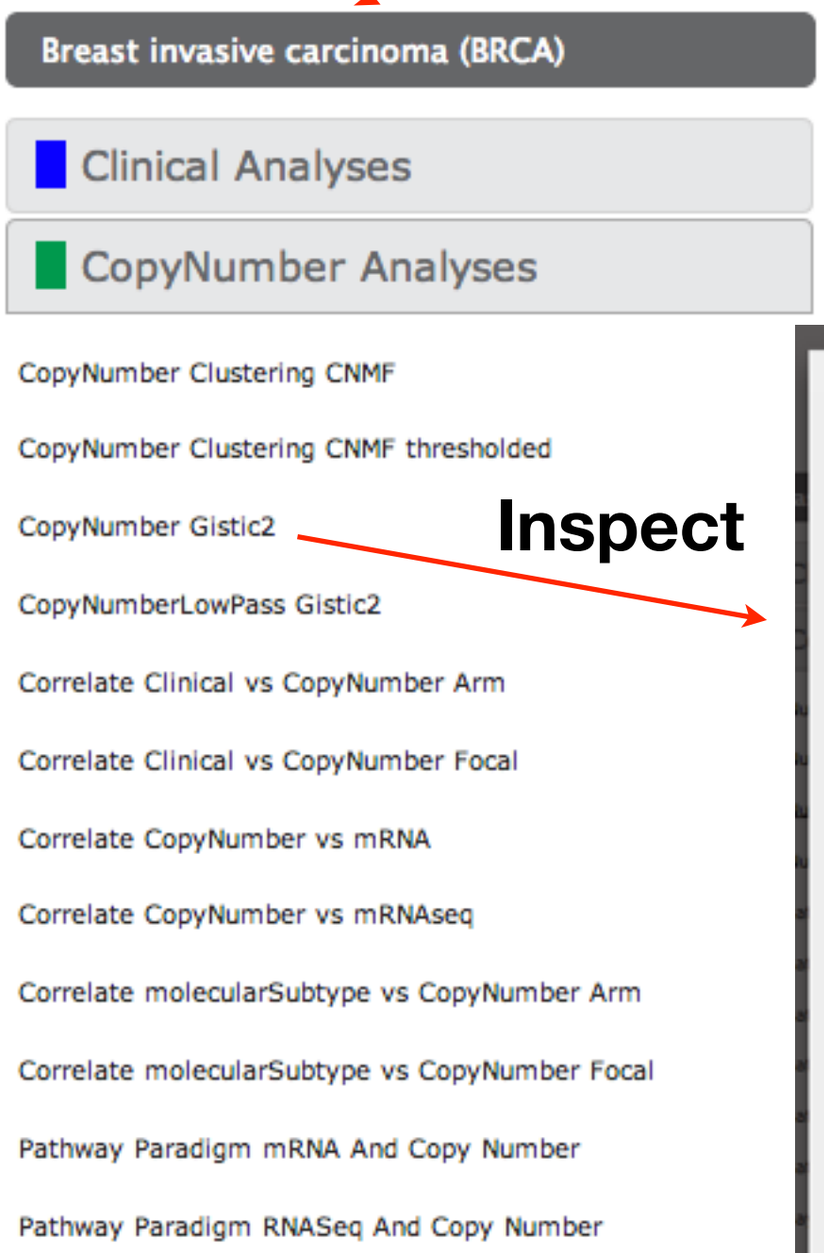
Simplifed Portal Access

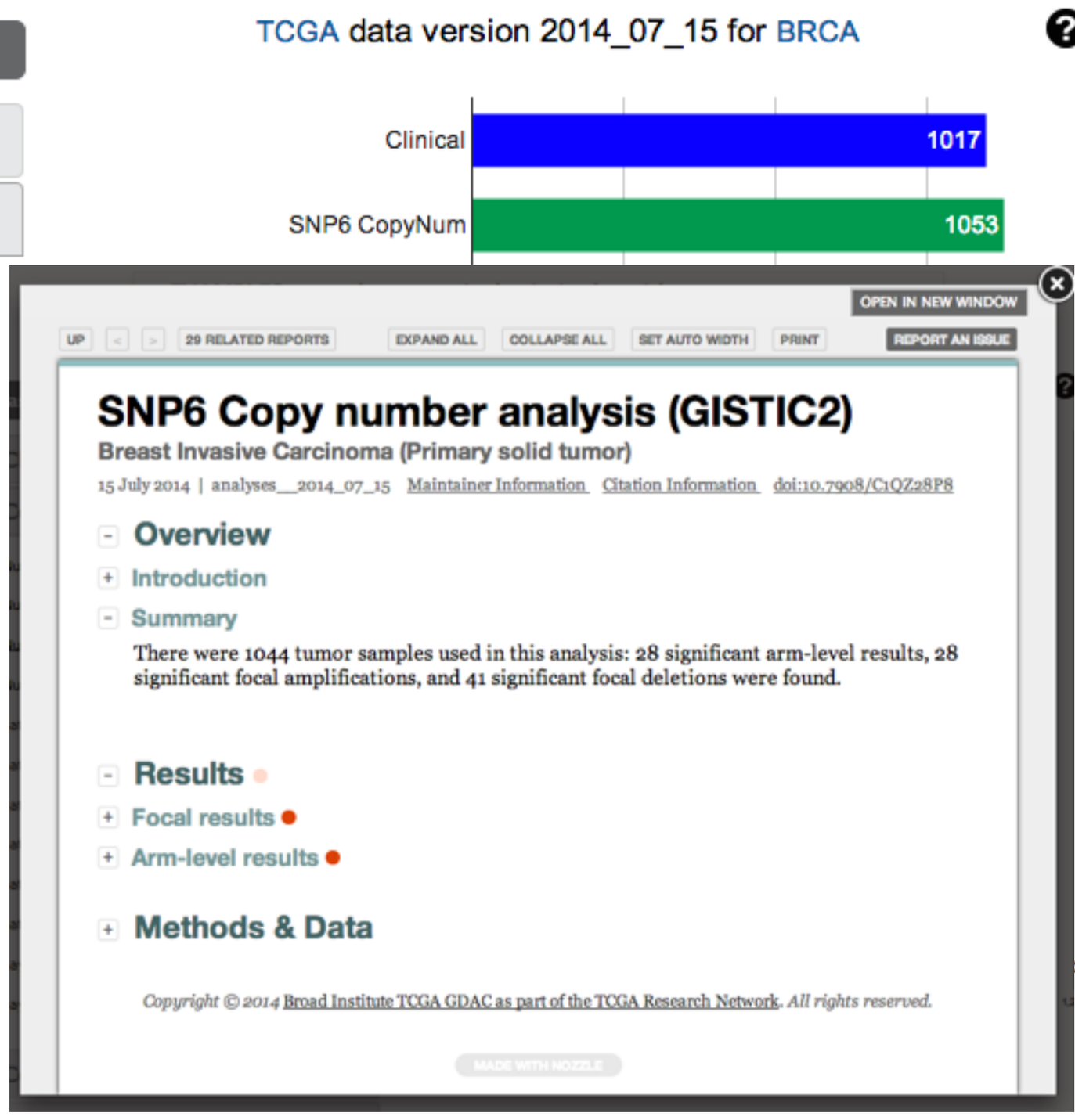
~1500 Analyses (reports) per run
Find your favorite in 2 clicks

Choose Cohort

Then DataType

Inspect





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

Choose Cohort

Thyroid carcinoma (THCA)

- Clinical Analyses
- CopyNumber Anal
- Correlations Analy
- Methylation Analys
- miRseq Analyses
- mRNA Analyses
- mRNAseq Analyse
- Mutation Analyses
- Pathway Analyses
- RPPA Analyses

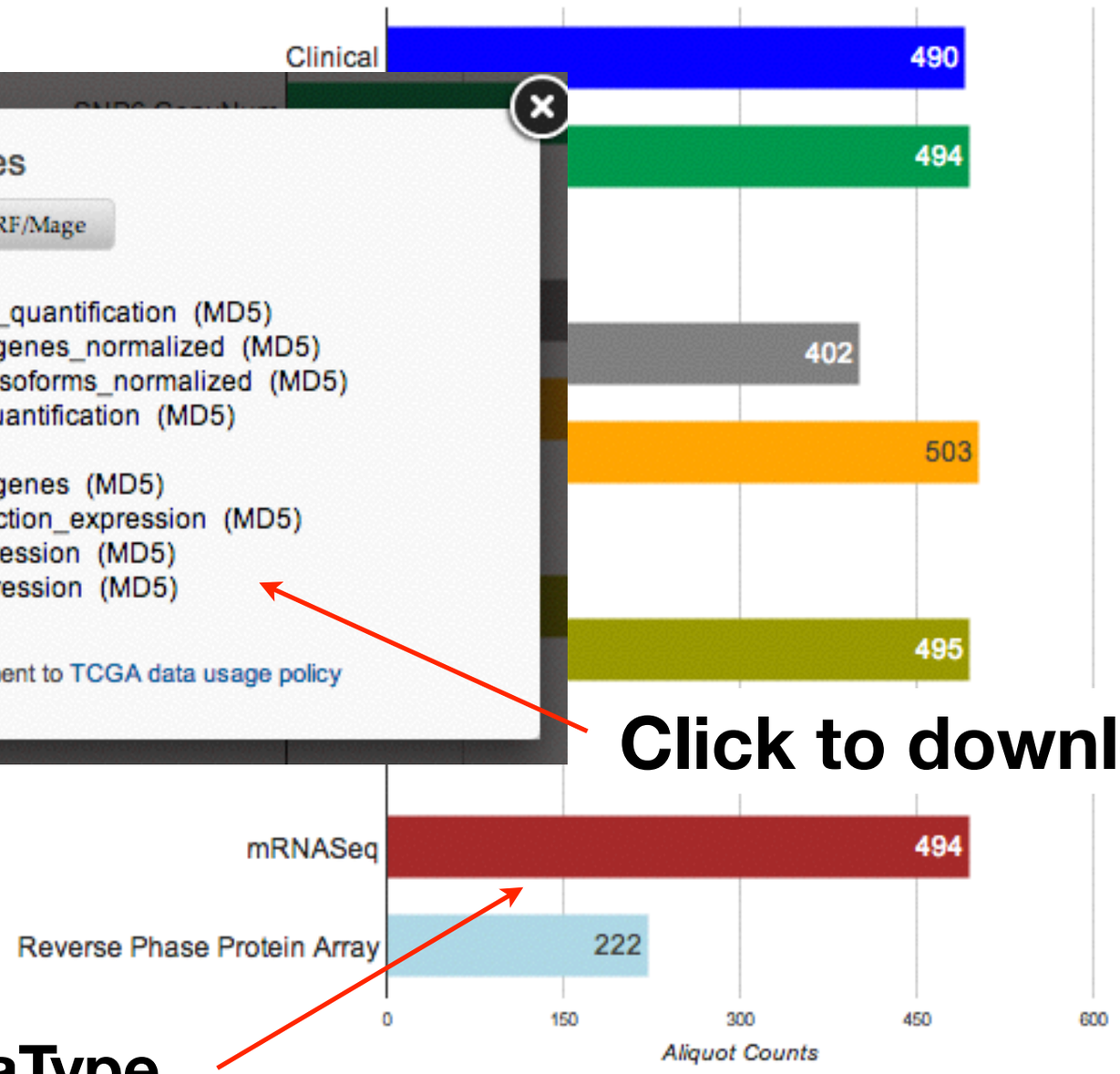
TCGA data version 2014_07_15 for THCA

THCA mRNASeq Archives

Primary | Auxiliary | SDRF/Mage

illuminahiseq_rnaseqv2-junction_quantification (MD5)
illuminahiseq_rnaseqv2-RSEM_genes_normalized (MD5)
illuminahiseq_rnaseqv2-RSEM_isoforms_normalized (MD5)
illuminahiseq_rnaseqv2-exon_quantification (MD5)
mRNAseq_Preprocess (MD5)
illuminahiseq_rnaseqv2-RSEM_genes (MD5)
illuminahiseq_rnaseq-splice_junction_expression (MD5)
illuminahiseq_rnaseq-exon_expression (MD5)
illuminahiseq_rnaseq-gene_expression (MD5)

Downloading data constitutes agreement to TCGA data usage policy



Click to download

Then DataType

API-Powered : 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

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

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

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
- Copiously flexible, documented and tested
- BSD-style open source license

https://confluence.broadinstitute.org/display/GDAC/fbget

FireBrowseR

R bindings developed by a Ph.D candidate in Germany
https://github.com/mariodeng/FirebrowseR

fbget : low level interface

```
python>  import firebrowse
python>  print  firebrowse.Samples().mRNASeq(gene="egfr", cohort="ucs")
{
  "mRNASeq": [
    {
      "cohort": "UCS",
      "expression_log2": 7.06162500904694,
      "gene": "EGFR",
      "geneID": 1956,
      "protocol": "RSEM",
      "sample_type": "TP",
      "tcga_participant_barcode": "TCGA-QN-A5NN",
      "z-score": -0.598993525060403

    },
    ...
```

4 classes, one per API category:
Samples, Analyses, Archives, Metadata

N methods per class, matching RESTful API; each defaults to returning 1 page, in JSON

fbget : high level interface

```
python>  import fbget
python>  print fbget.mrnaseq("egfr", cohort="ucs")

tcga_participant_barcode        gene    expression_log2 z-score cohort
TCGA-QN-A5NN    EGFR    7.06162500905   -0.59899352506  UCS     TP
TCGA-QM-A5NM    EGFR    8.16734387649   -0.298443593752 UCS     TP
TCGA-NG-A4VW    EGFR    8.93092623547   0.0932667888031 UCS     TP
```

- Simpler, e.g. objects do not need to be instantiated
- Intuitive defaults for common bioinformatic use cases
- Transparently iterates:
  - ✓ To retrieve all pages of results in 1 call
  - ✓ In TSV format

fbget : UNIX CLI interface

```
linux%   fbget mrnaseq egfr cohort=ucs

tcga_participant_barcode        gene    expression_log2 z-score cohort
TCGA-QN-A5NN    EGFR    7.06162500905   -0.59899352506  UCS     TP
TCGA-QM-A5NM    EGFR    8.16734387649   -0.298443593752 UCS     TP
TCGA-NG-A4VW    EGFR    8.93092623547   0.0932667888031 UCS     TP
```

Because sometimes even writing just a couple of lines of Python takes too long

Example: quickly list patients

All of TCGA

```
linux%  fbget patients

tcga_participant_barcode     date     cohort
TCGA-PK-A5H9     2015-04-02 00:00:00 ACC
TCGA-PA-A5YG     2015-04-02 00:00:00 ACC
TCGA-OR-A5JD     2015-04-02 00:00:00 ACC
TCGA-P6-A5OF     2015-04-02 00:00:00 ACC
TCGA-P6-A5OG     2015-04-02 00:00:00 ACC
```

Or just GBM

```
linux%  fbget patients  cohort=gbm

tcga_participant_barcode     date     cohort
TCGA-19-4065     2015-04-02 00:00:00 GBM
TCGA-81-5911     2015-04-02 00:00:00 GBM
TCGA-81-5910     2015-04-02 00:00:00 GBM
TCGA-12-1089     2015-04-02 00:00:00 GBM
```

This can be enhanced to yield platform data matrix, like AWG freeze list

fbget Documentation

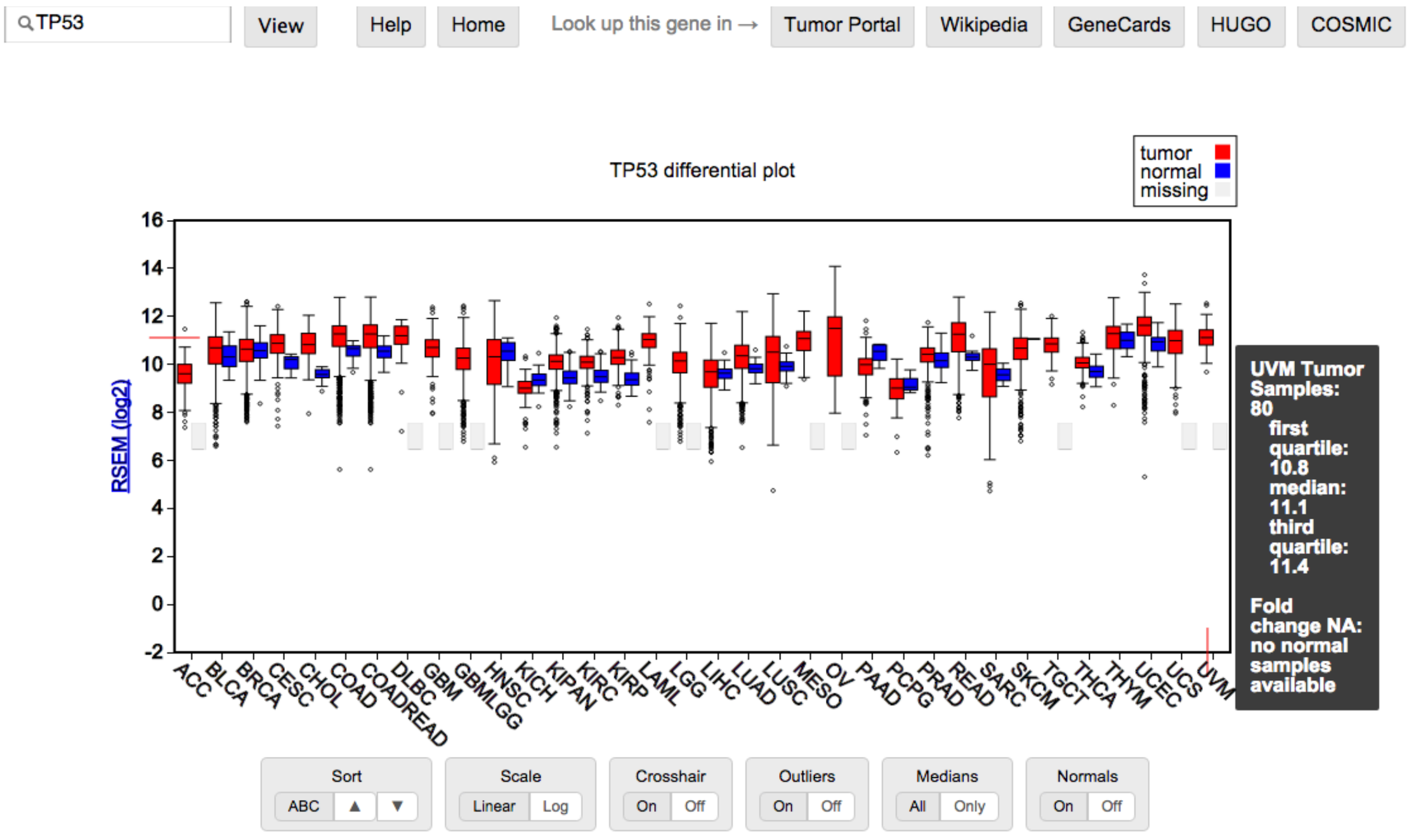
- Website
- fbget —examples
- Python help

Docs for almost all class methods and functions can also be obtained by invoking the function with zero arguments.

```
python>  fbget.mrnaseq()

mrnaseq() call has missing/None arg value(s), need at least one of: gene OR barcode
Help on function mrnaseq in module fbget:

mrnaseq(gene=None, barcode=None, **kwargs)

    High level wrapper for the FireBrowse Samples.mRNASeq method.
    By default it returns ALL pages of data, in TSV format. . .
```

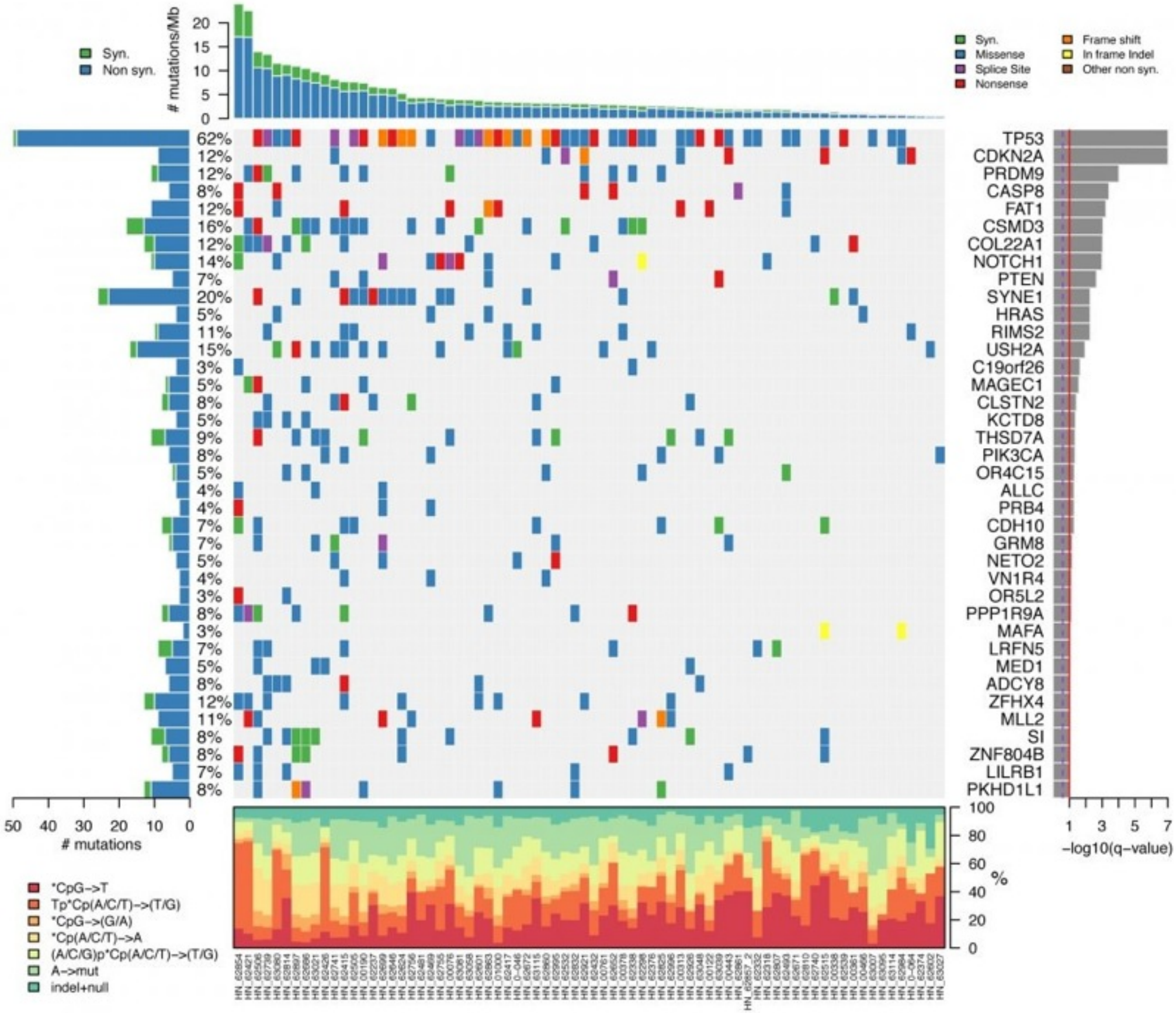
Better than an inscrutable stack trace, don't you think?

New Visualization: firebrowse.org/viewGene.html

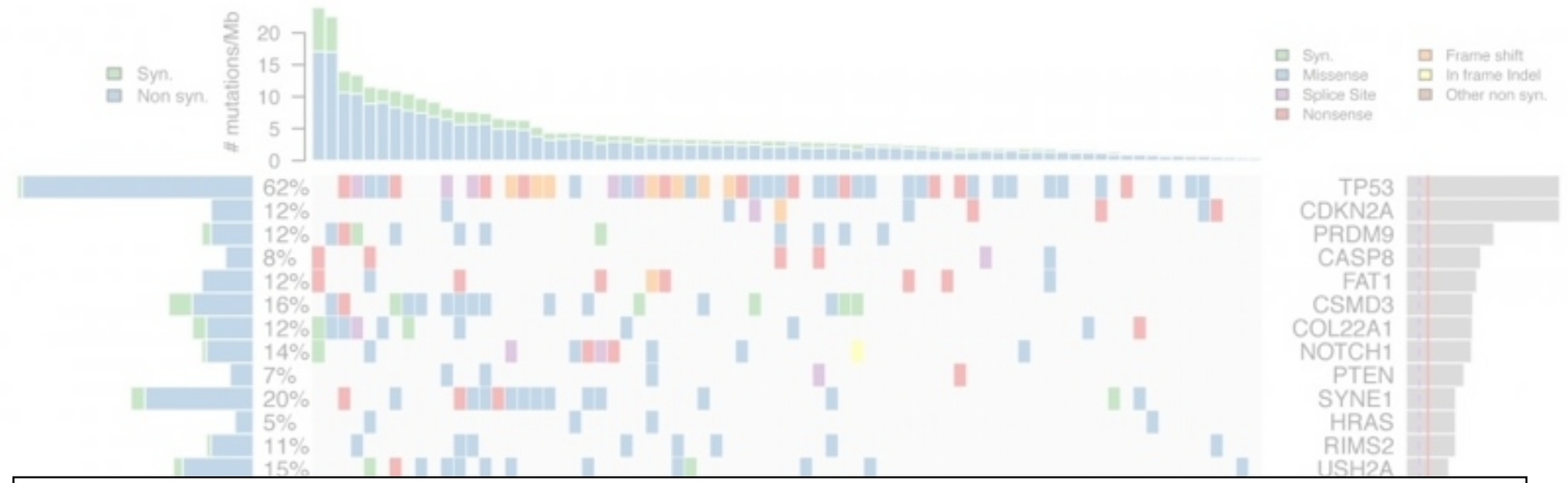


Built on top of the FireBrowse API, lets one quickly inspect mRNASeq expression levels for a selected gene, across all cohorts.

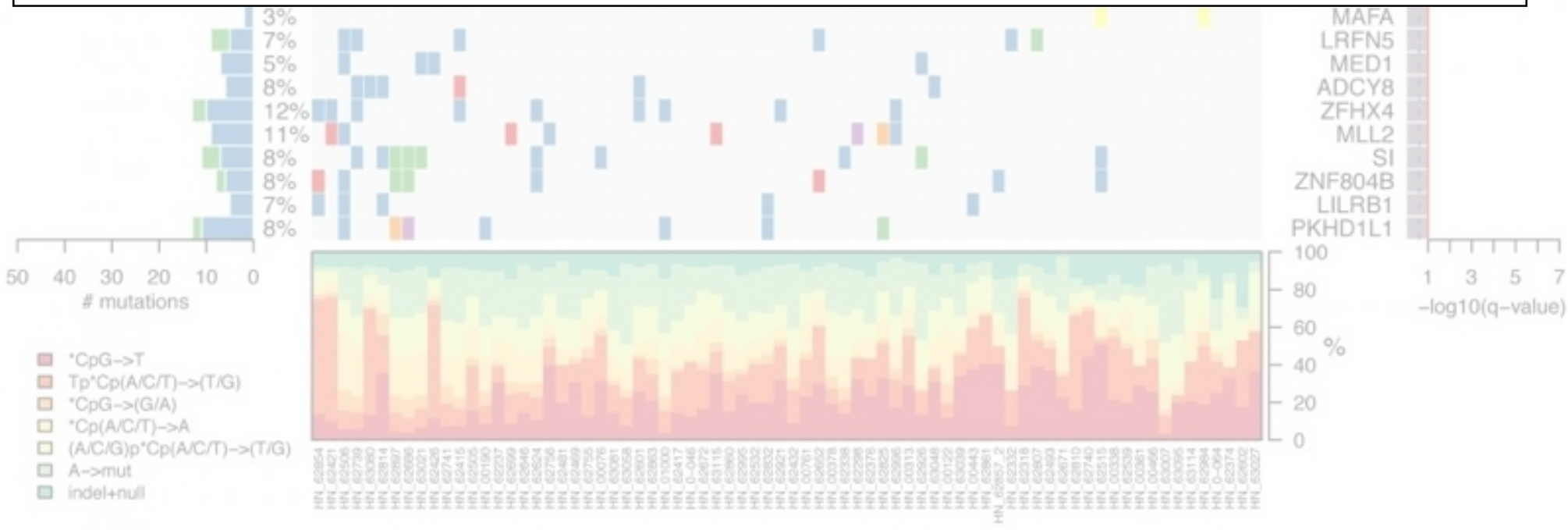
New Visualization: iCoMut

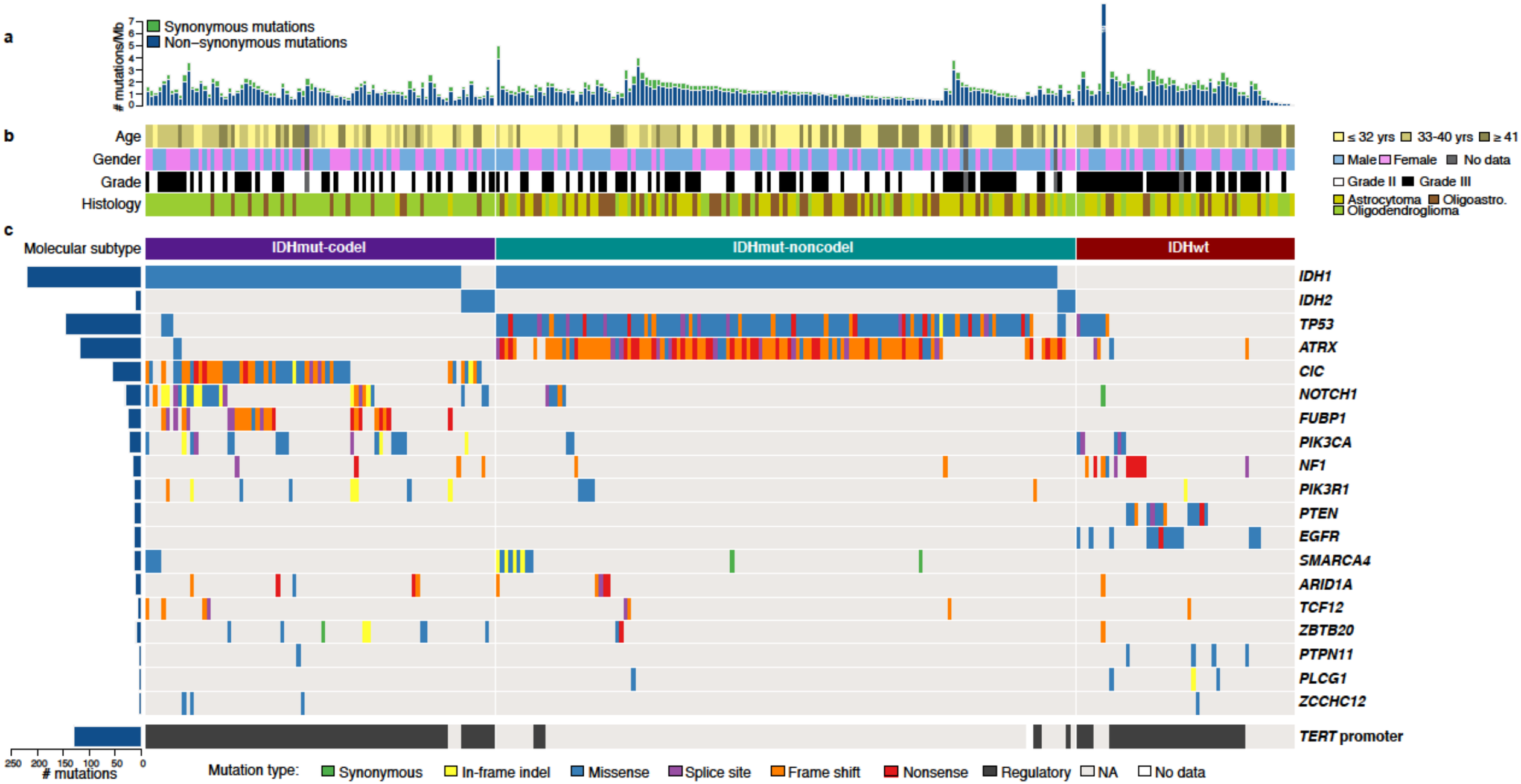


New Visualization: iCoMut



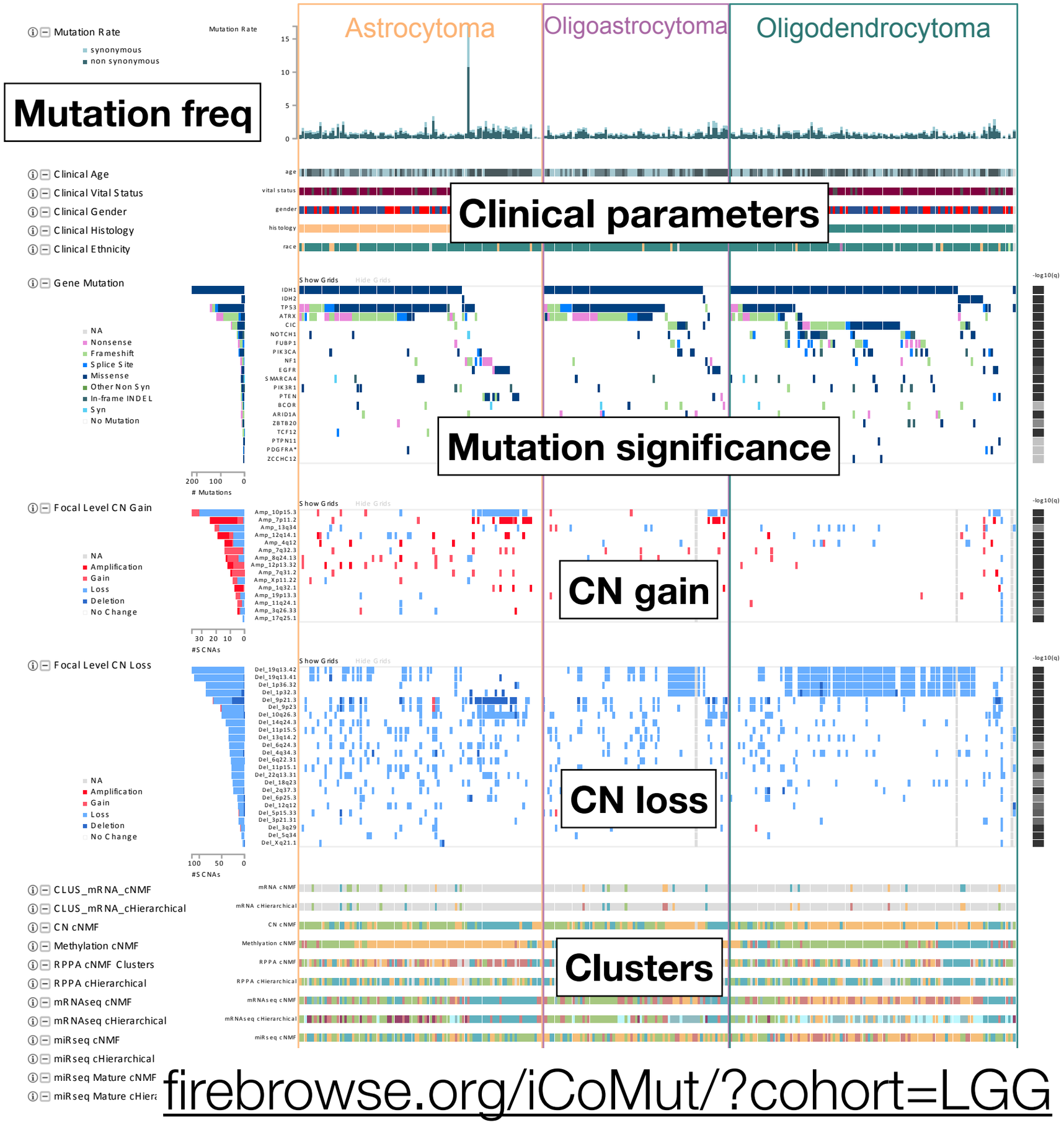
Introduced by N. Stransky (*The Mutational Landscape of Head and Neck Squamous Cell Carcinoma*. *Science,* 2011), CoMut figures have become common in TCGA research. Within a single graphic they provide a *comprehensive analysis profile*, enabling the reader to quickly infer relationships between co-occurring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





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

Figure courtesy of Jaegil Kim, Broad Institute



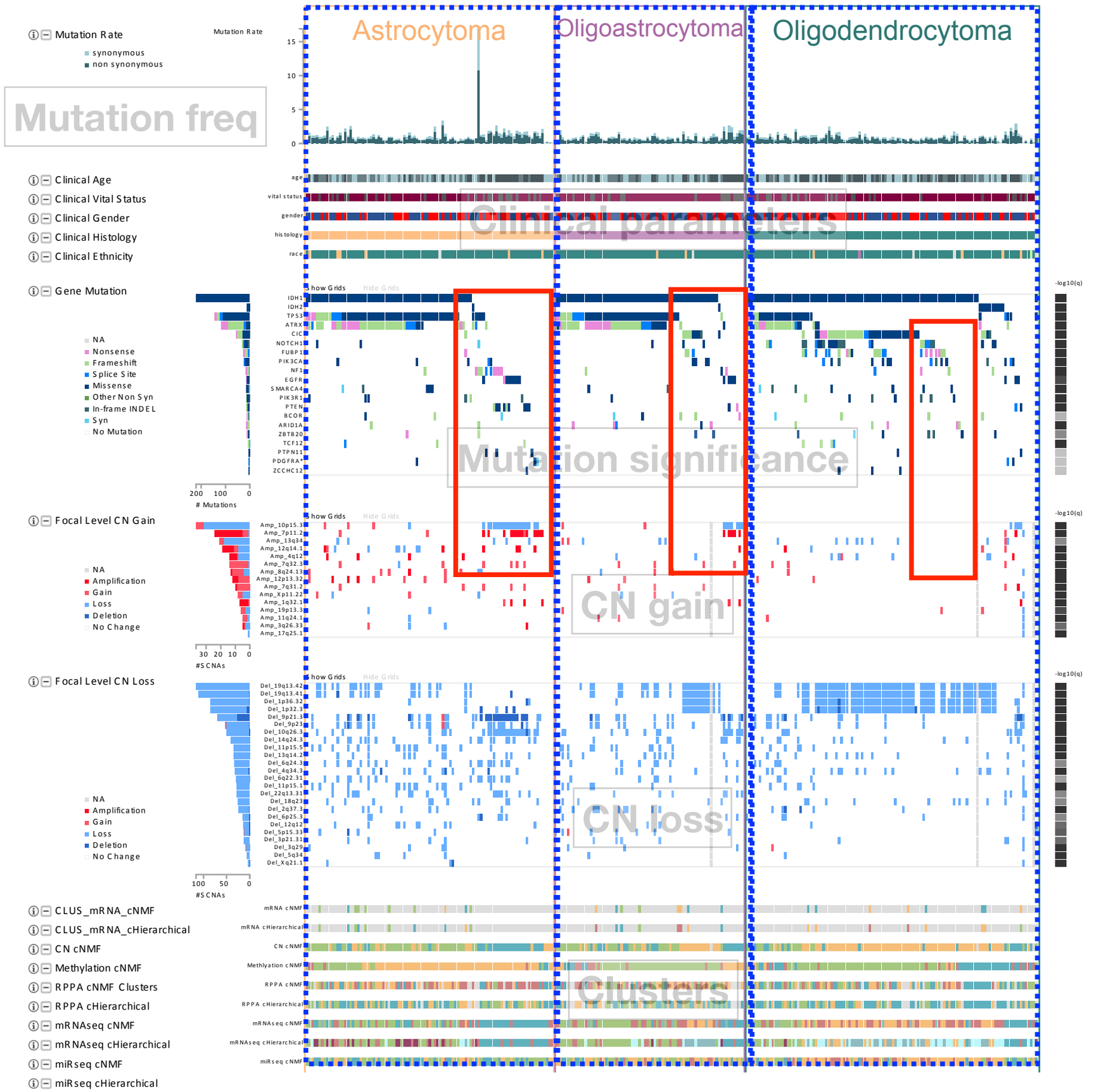
firebrowse.org/iCoMut/?cohort=LGG

Here we show the TCGA LGG cohort: sorted first by clinical histology, then gene (descending order of mutation count). It is quickly apparent that copy-number changes differ when IDH1/2, TP53, and ATRX mutations drop off.

firebrowse.org/iCoMut/?cohort=LGG

Here we show the TCGA LGG cohort: sorted first by clinical histology, then gene (descending order of mutation count). It is quickly apparent that copy-number changes differ when IDH1/2, TP53, and ATRX mutations drop off.

The LGG subtypes are also very clear

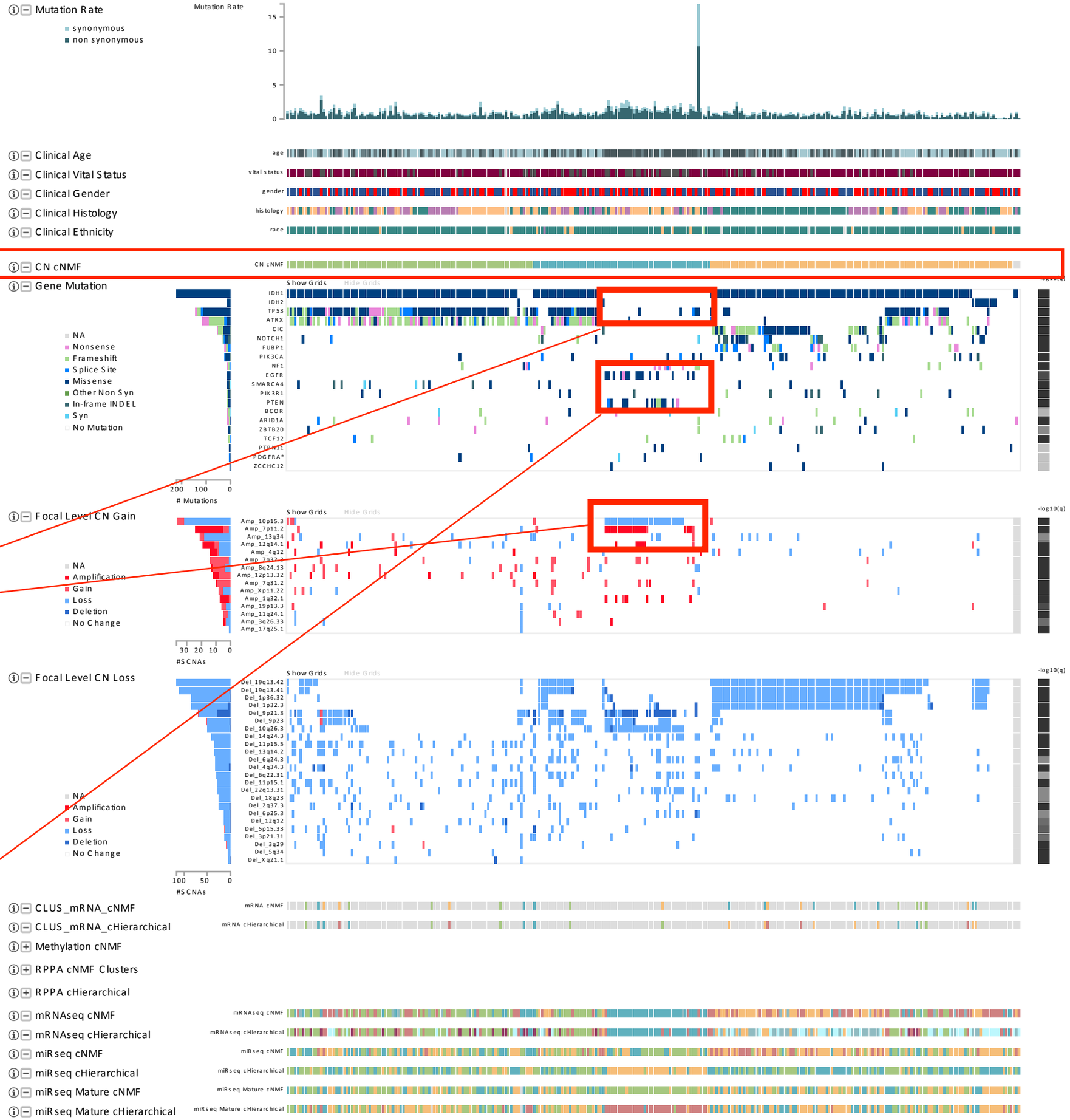


firebrowse.org/iCoMut/?cohort=LGG

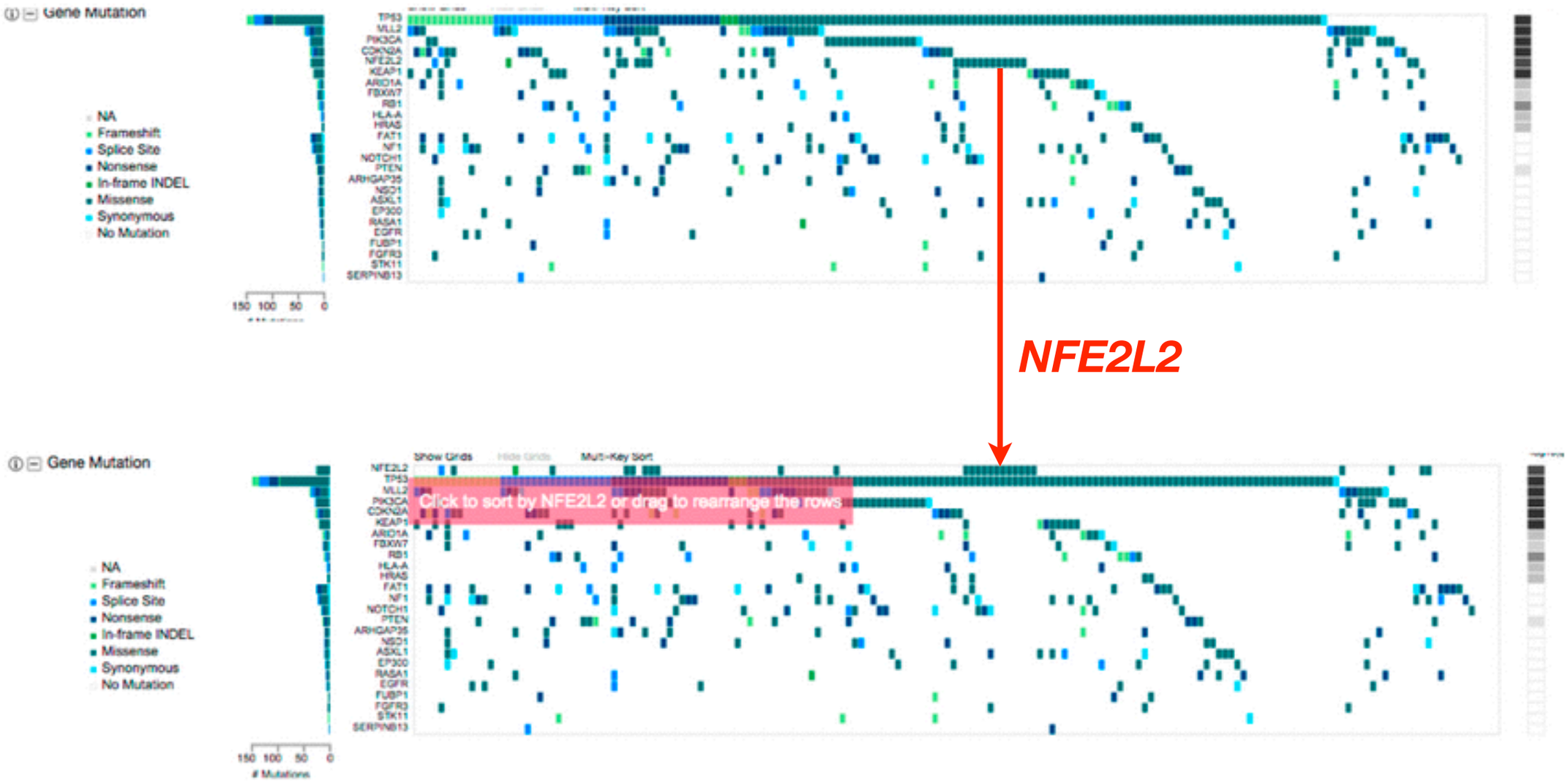
Here we've re-sorted by CNMF copy-number clustering, *and dragged it from bottom of graphic to top*, just above mutation panel

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.



Drag and drop the row names to rearrange the row order



NFE2L2

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

32 of 38 disease cohorts ready for inspection
(other 6 have no mutation data yet)

Expected out of beta by end of summer

Further Work

- Manage richness of information
- Magnifying glass zoom would help
- Performance
- Data import & export

viewGene

iCoMut

Integrated directly into firebrowse.org

Fin