GeneSpot

A portal for interactive gene-centric exploration of The Cancer Genome Atlas

Brady Bernard & Hector Rovira

Shmulevich and Zhang TCGA GDAC

The Cancer Genome Atlas

Institute for Systems Biology

Revolutionizing Science. Enhancing Life.

Motivation

- For a given gene, for any TCGA tumor type:
  - What is the mutation profile?
  - Are there significant copy number aberrations?
  - What are the data-derived statistical associations?
  - What would a plot of Gene A and Gene B look like?

Motivation

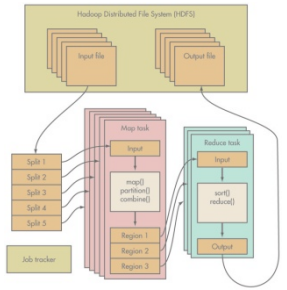
- For a given gene, for any TCGA tumor type:
  - What is the mutation profile?
  - Are there significant copy number aberrations?
  - What are the data-derived statistical associations?
  - What would a plot of Gene A and Gene B look like?
- Such gene-centric questions are not trivial in practice
  - Data repositories are largely organized in a sample-centric or tumor-centric manner

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

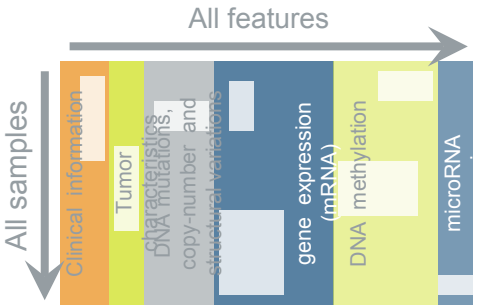
Typical Workflow

- Download all data
  - TCGA Data Portal or Broad Firehose

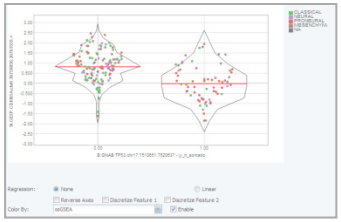
- Parse and process data
  - e.g., parse MAGE-TAB SDRF to determine Level_3 file mappings, relate features with genomic coordinates to genes



- Merge all data and extract features associated with gene(s) of interest
  - e.g., retain all TP53 associated columns

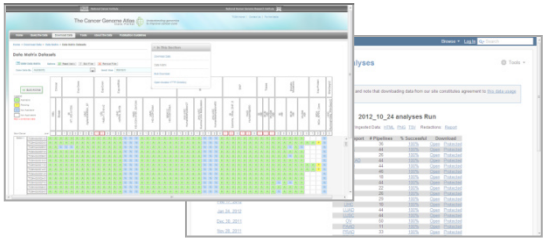


- Analyze and create figures
  - R, Excel

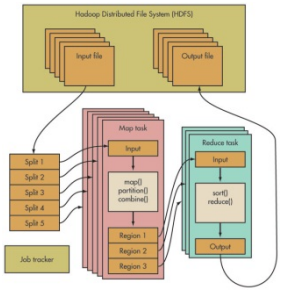


Typical Workflow

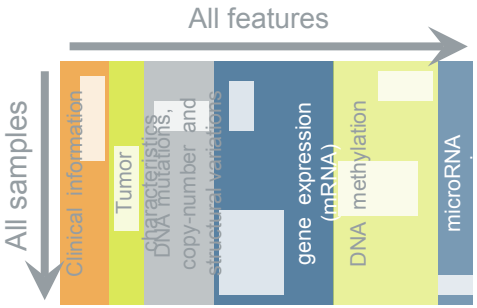
- Download all data
  - TCGA Data Portal or Broad Firehose



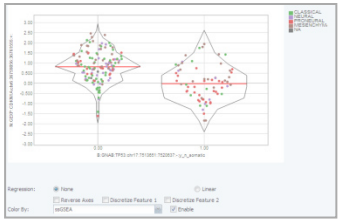
- Parse and process data
  - e.g., parse MAGE-TAB SDRF to determine Level_3 file mappings, relate features with genomic coordinates to genes



- Merge all data and extract features associated with gene(s) of interest
  - e.g., retain all TP53 associated columns

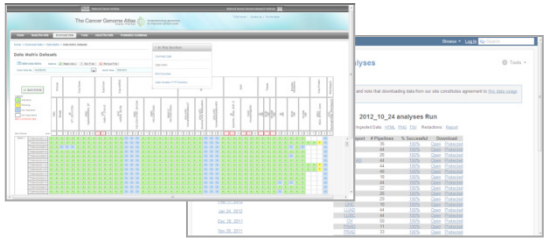


- Analyze and create figures
  - R, Excel

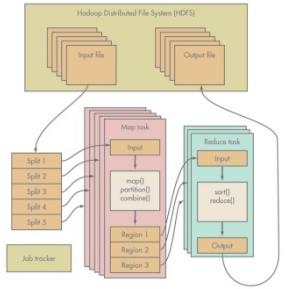


Typical Workflow

- Download all data
  - TCGA Data Portal or Broad Firehose

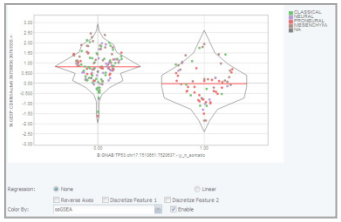


- Parse and process data
  - e.g., parse MAGE-TAB SDRF to determine Level_3 file mappings, relate features with genomic coordinates to genes



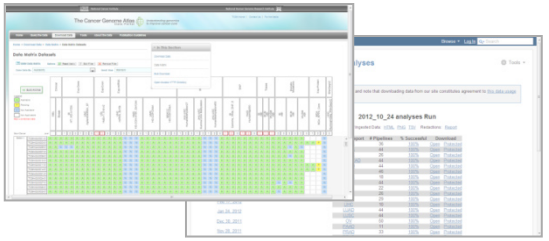
- Merge all data and extract features associated with gene(s) of interest
  - e.g., retain all TP53 associated columns

- Analyze and create figures
  - R, Excel

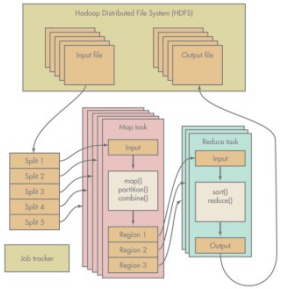


Typical Workflow

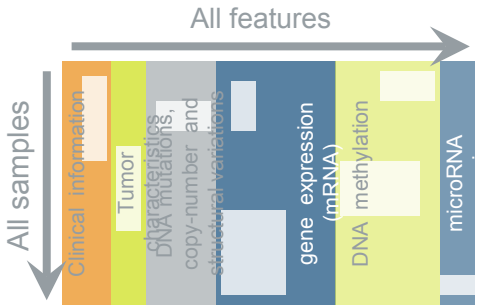
- Download all data
  - TCGA Data Portal or Broad Firehose



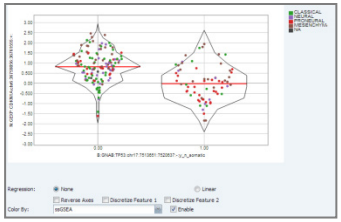
- Parse and process data
  - e.g., parse MAGE-TAB SDRF to determine Level_3 file mappings, relate features with genomic coordinates to genes



- Merge all data and extract features associated with gene(s) of interest
  - e.g., retain all TP53 associated columns

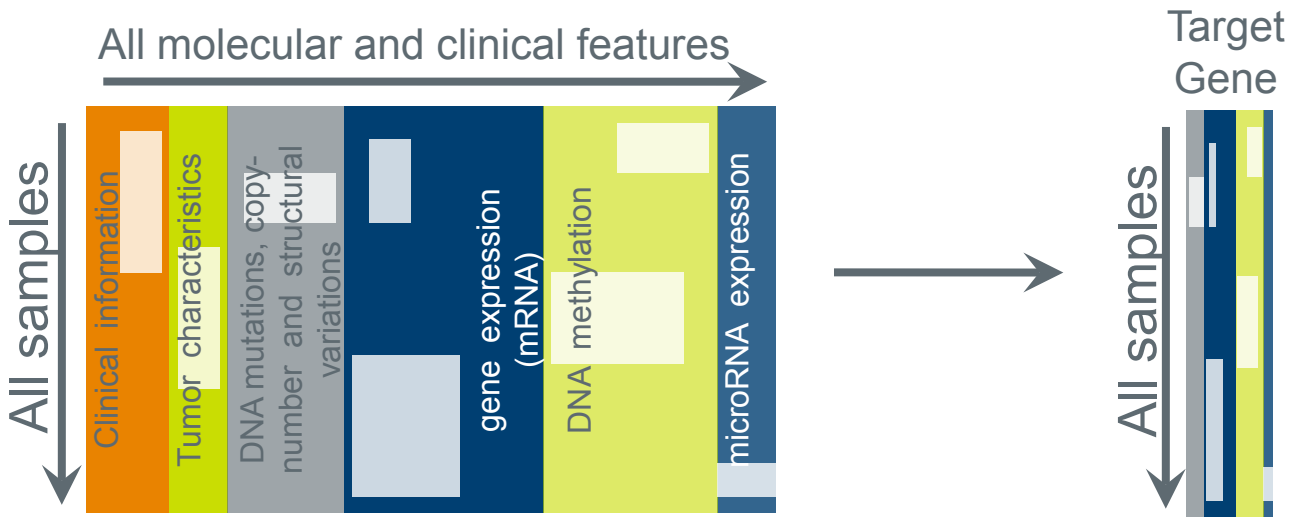


- Analyze and create figures
  - R, Excel



Challenges

- Data required for gene-centric analysis
  - ~ 500k data points per biological sample
  - ~ 10k samples across all tumor types
  - ~ 5 billion data points
  - ~ 200 Gb data
- Significant time, resources, and expertise required
- Only thousands of data points needed for gene-centric analysis



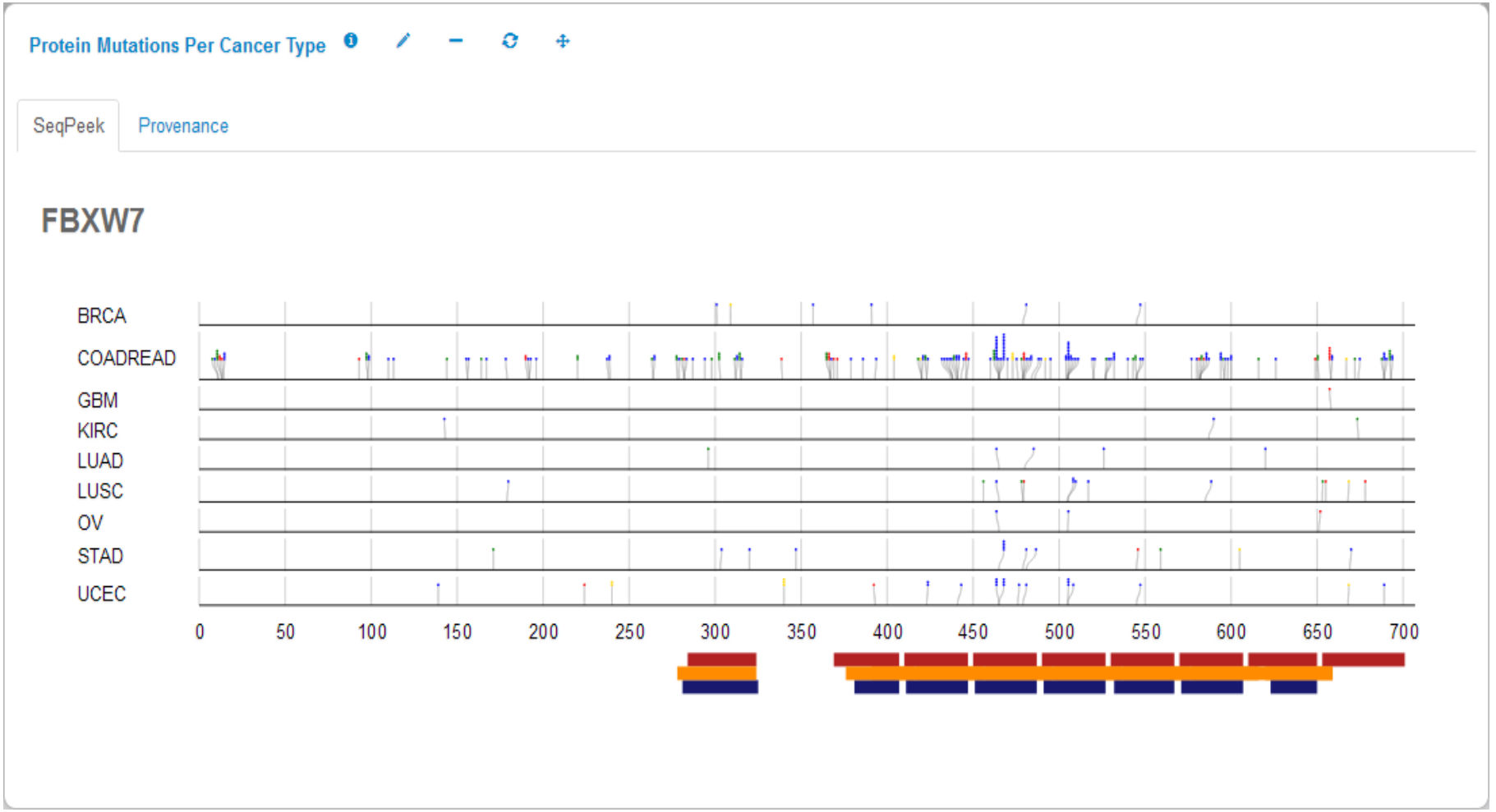
GeneSpot Approach

- Interactive Web Portal
  - Gene or gene sets are specified and explored
  - No need to download data or install software
- Controllable Canvas
  - Numerous gene-centric views available
  - Views can be moved, expanded, minimized, removed from the canvas
- Sessions
  - The state of the exploration can be saved and shared, enabling collaboration and retrieval of several gene-centric views
- Direct Data Access
  - Data table downloads allow direct gene-centric access to mirrored data repositories

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Example Views

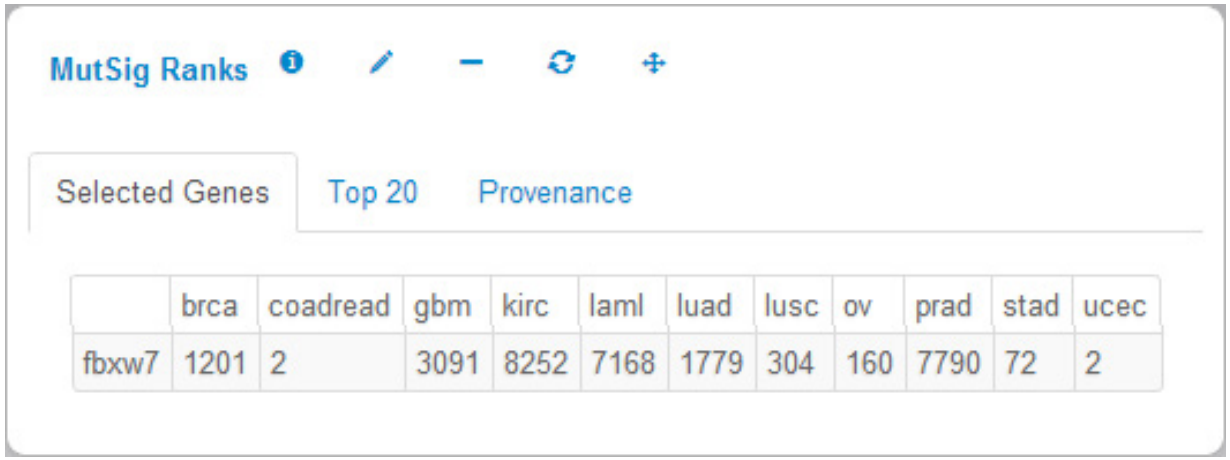
FBXW7 Mutations



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Example Views

FBXW7 Mutations

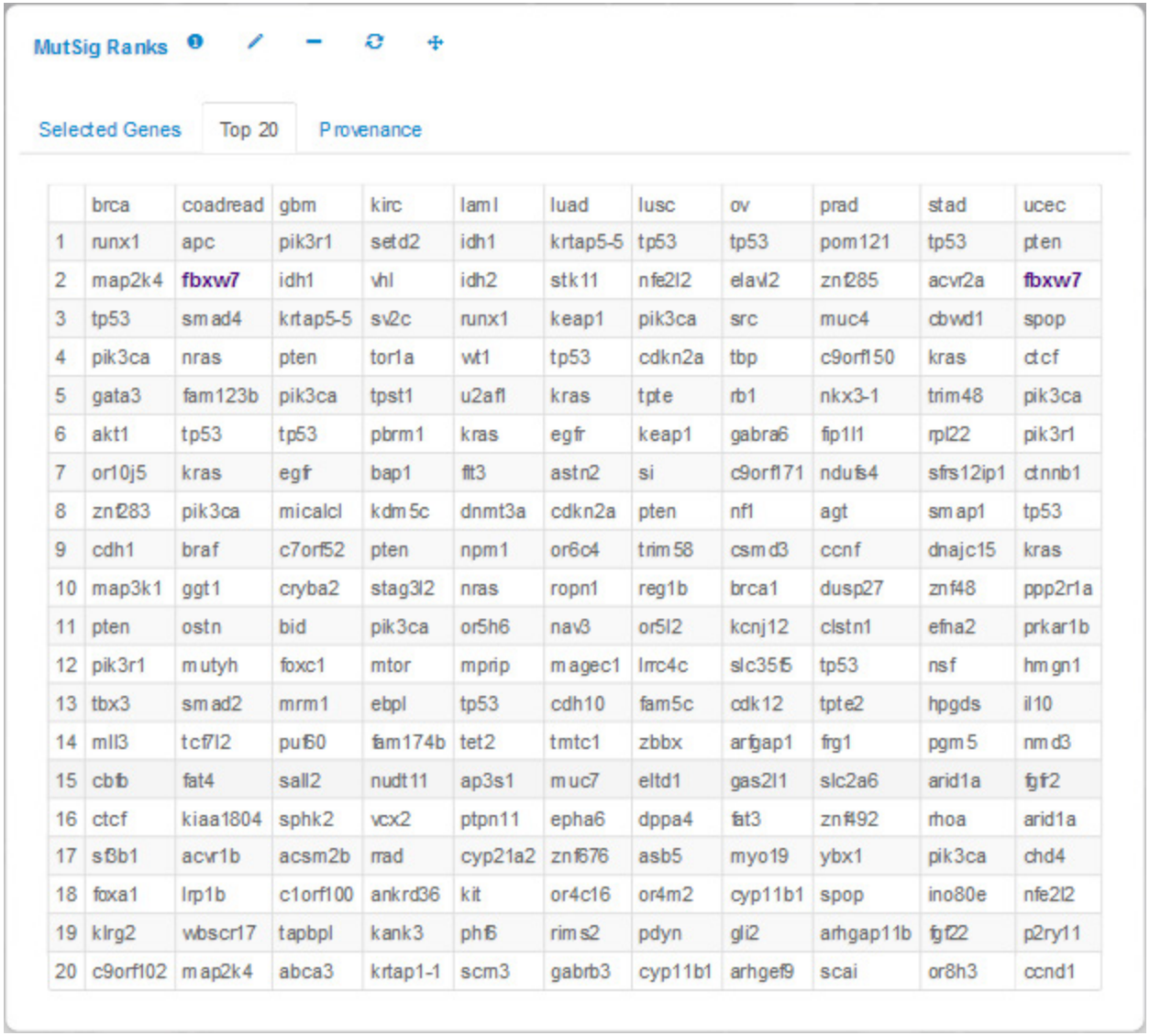


	brca	coadread	gbm	kirc	laml	luad	lusc	ov	prad	stad	ucec
fbxw7	1201	2	3091	8252	7168	1779	304	160	7790	72	2

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Example Views

MutSig Top 20



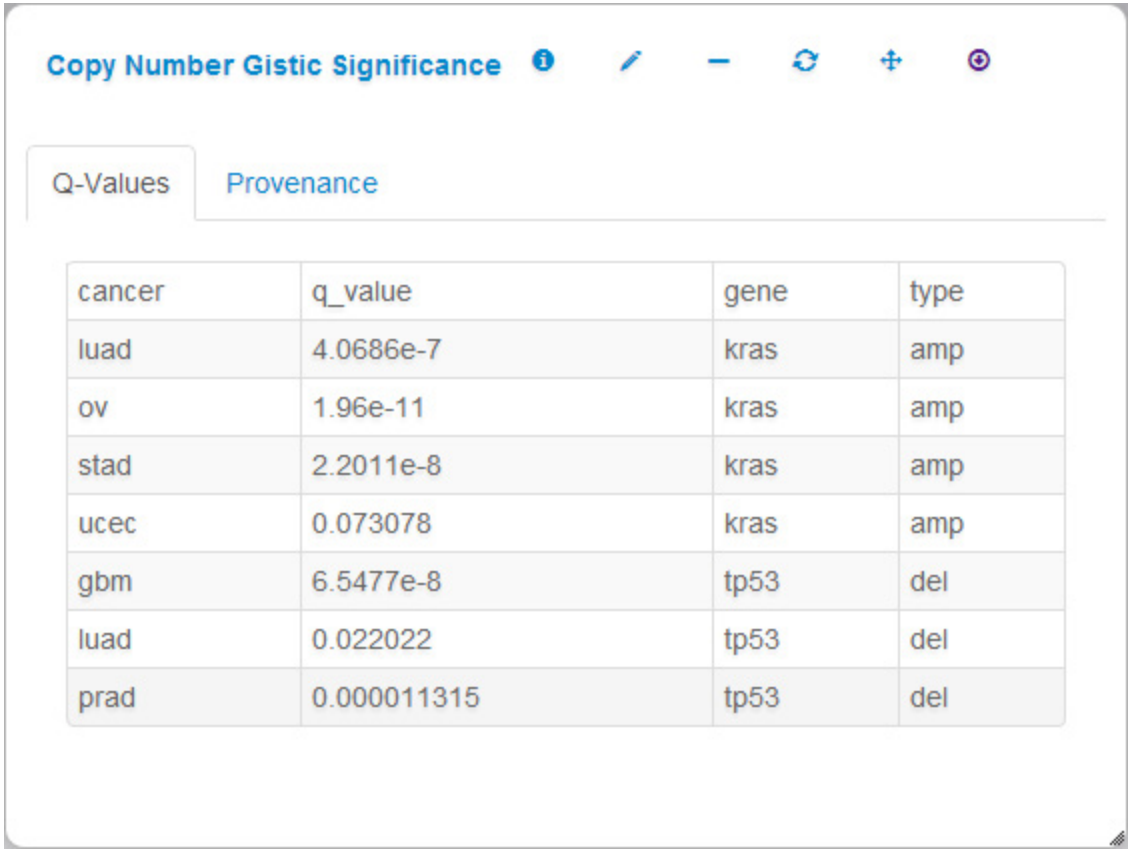
MutSig Ranks

Selected Genes | Top 20 | Provenance

	brca	coadread	gbm	kirc	laml	luad	lusc	ov	prad	stad	ucec
1	runx1	apc	pik3r1	setd2	idh1	krtap5-5	tp53	tp53	pom121	tp53	pten
2	map2k4	fbxw7	idh1	vhl	idh2	stk11	nfe2l2	elavl2	znf285	acvr2a	fbxw7
3	tp53	smad4	krtap5-5	sv2c	runx1	keap1	pik3ca	src	muc4	cbwd1	spop
4	pik3ca	nras	pten	tor1a	wt1	tp53	cdkn2a	tbp	c9orf150	kras	ctcf
5	gata3	fam123b	pik3ca	tpst1	u2af1	kras	tpte	rb1	nkx3-1	trim48	pik3ca
6	akt1	tp53	tp53	pbrm1	kras	egfr	keap1	gabra6	fip1l1	rpl22	pik3r1
7	or10j5	kras	egfr	bap1	flt3	astn2	si	c9orf171	ndufs4	sfrs12ip1	ctnnb1
8	znf283	pik3ca	micalcl	kdm5c	dnmt3a	cdkn2a	pten	nf1	agt	smap1	tp53
9	cdh1	braf	c7orf52	pten	npm1	or6c4	trim58	csmd3	ccnf	dnajc15	kras
10	map3k1	ggt1	cryba2	stag3l2	nras	ropn1	reg1b	brca1	dusp27	znf48	ppp2r1a
11	pten	ostn	bid	pik3ca	or5h6	nav3	or5l2	kcnj12	clstn1	efna2	prkar1b
12	pik3r1	mutyh	foxc1	mtor	mprip	magec1	lrrc4c	slc35f5	tp53	nsf	hmgn1
13	tbx3	smad2	mrm1	ebpl	tp53	cdh10	fam5c	cdk12	tpte2	hpgds	il10
14	mll3	tcf7l2	puf60	fam174b	tet2	tmtc1	zbbx	arfgap1	frg1	pgm5	nmd3
15	cbfb	fat4	sall2	nudt11	ap3s1	muc7	eltd1	gas2l1	slc2a6	arid1a	fgf2
16	ctcf	kiaa1804	sphk2	vcx2	ptpn11	epha6	dppa4	fat3	znf492	rhoa	arid1a
17	sf3b1	acvr1b	acsm2b	rrad	cyp21a2	znf676	asb5	myo19	ybx1	pik3ca	chd4
18	foxa1	lrp1b	c1orf100	ankrd36	kit	or4c16	or4m2	cyp11b1	spop	ino80e	nfe2l2
19	klrg2	wbscr17	tapbpl	kank3	phf6	rims2	pdyn	gli2	arhgap11b	fgf22	p2ry11
20	c9orf102	map2k4	abca3	krtap1-1	scrn3	gabrb3	cyp11b1	arhgef9	scai	or8h3	ccnd1

Example Views

Significant copy number aberrations

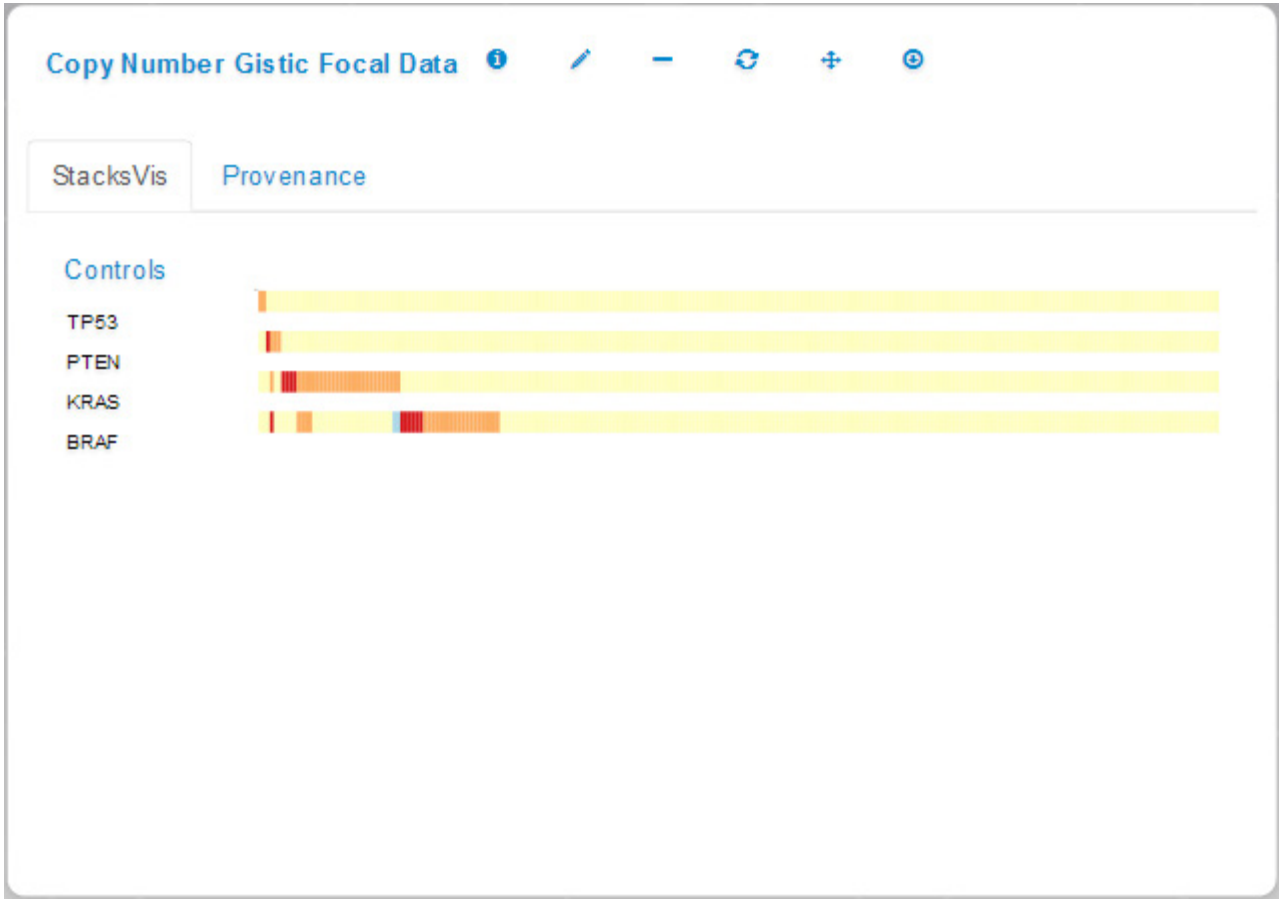


cancer	q_value	gene	type
luad	4.0686e-7	kras	amp
ov	1.96e-11	kras	amp
stad	2.2011e-8	kras	amp
ucec	0.073078	kras	amp
gbm	6.5477e-8	tp53	del
luad	0.022022	tp53	del
prad	0.000011315	tp53	del

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Example Views

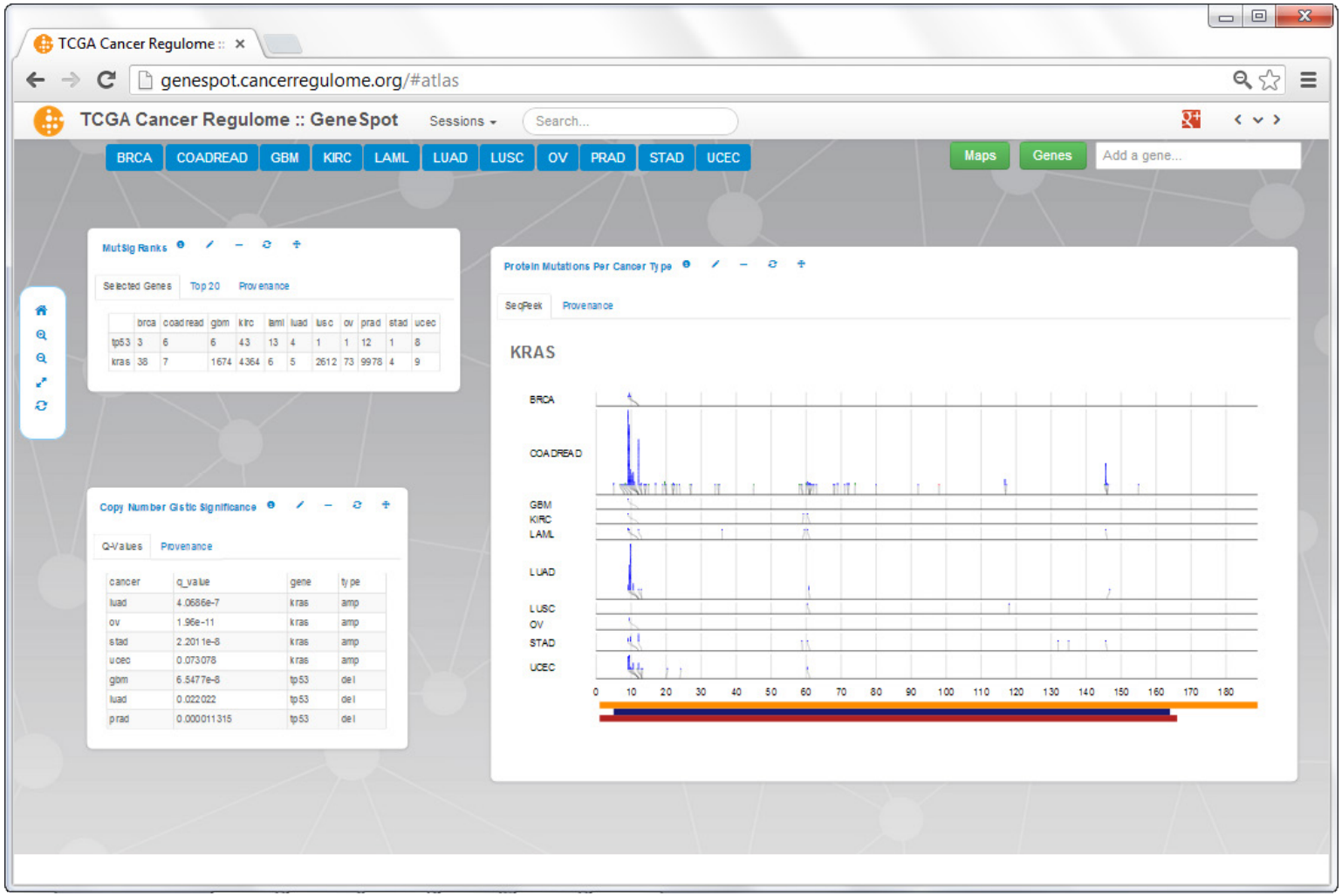
Focal copy Number



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

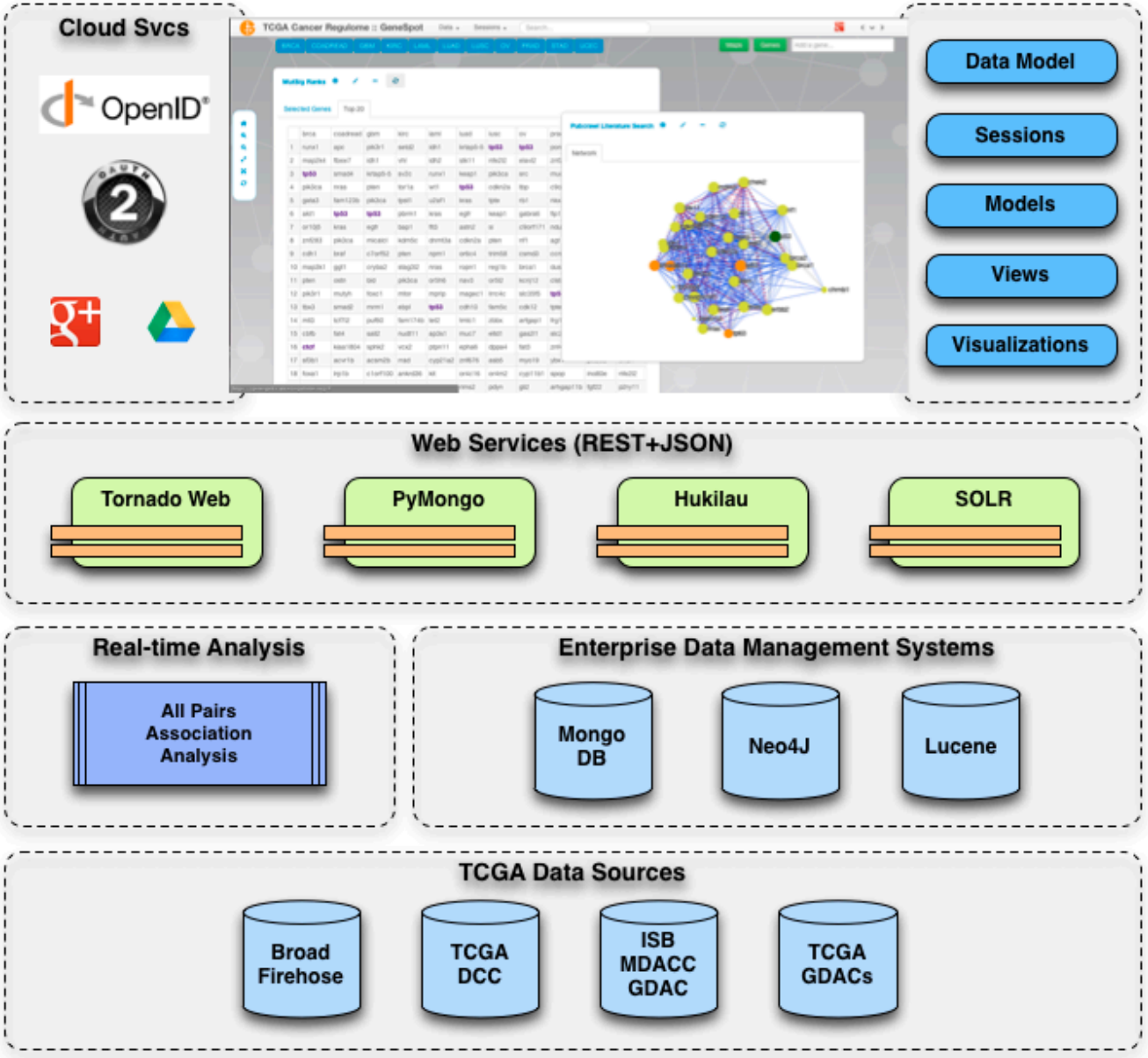
Demo

http://genespot.org



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Software Architecture



Future Directions & Integration

- Additional views
  - Integration with other analyses and views developed by TCGA community
- Role of target gene(s) in context of pathways
- Further integration with Google cloud services
- Provide deep links to share URLs

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Acknowledgements

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ilya Shmulevich
Kalle Leinonen
Richard Kreisberg
Andrea Eakin
Sheila Reynolds
Jake Lin
Roger Kramer
Lisa Iype
Ryan Bressler
Vesteinn Thorsson

National Cancer Institute

Award Number U24CA143835

The Cancer Genome Atlas

FIREHOSE
Broad GDAC

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Wei Zhang
Da Yang
Yuexin Liu

Memorial Sloan-Kettering Cancer Center
The Best Cancer Care. Anywhere.

cBio Cancer Genomics Portal

The Genome Institute at Washington University

http://genespot.org